On the Factors Shaping Numerical Real-World Estimation

Barbara K. Kreis

Inaugural Dissertation
Submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

Thesis Defense:
19.12.2025

Supervisors:
Dr. Julia Groß
Prof. Dr. Thorsten Pachur

Dean of the School of Social Sciences:
Dr. Julian Dierkes

Thesis Reviewers:
Prof. Dr. Arndt Bröder
Prof. Dr. Edgar Erdfelder

For all those who accompanied me on this journey

Contents

Summary

Numerical estimation is a fundamental cognitive skill supporting numerous everyday decisions, from planning travel time and estimating food quantities to judging financial costs. Yet despite extensive research, our understanding of numerical real-world estimation remains fragmented, as foundational mechanisms, phenomena of altered estimation, and applied questions are often studied separately. This dissertation addresses this gap by examining numerical real-world estimation through an integrative lens of the factors that shape it. With Manuscript I, I broaden the understanding of the mental resources that constitute the foundation of numerical estimation. Whereas prior work focused on the role of domain knowledge, I demonstrate that symbolic-number mapping, a basic numeric ability typically studied in numerical cognition, also underlies numerical estimation. In Manuscripts II and III, I deliberate the malleability of numerical estimation as a consequence of changes in the underlying resources, specifically in the knowledge base. In Manuscript II, I examine how estimation accuracy can be improved in real-world contexts to support informed decision-making by providing information that triggers knowledge updating. Specifically, I compare three different informational interventions—one based on cognitive research, two based on applied research—each targeting different aspects of knowledge. The results showed that the cognitive intervention, which provides actual values for estimated items, improved knowledge and thus estimation accuracy most comprehensively and supported informed decision-making equally well as the applied interventions. In Manuscript III, I examine hindsight bias—the tendency to recall prior estimates as more accurate after learning the actual values. Traditionally studied separately from numerical estimation research, I argue that it can be understood as a phenomenon of numerical estimation being shaped by changes in knowledge. Building on a co-authored study (Groß et al., 2023), I provide evidence that hindsight bias is not merely a superficial memory distortion, but the by-product of the same updating processes that also improve estimation accuracy. With this dissertation, I address the role of basic numeric abilities and domain knowledge in fundamentally shaping numerical real-world estimation. Integrating previously separate lines of research, I deepen our theoretical understanding while also providing guidance for supporting real-world estimation.

Manuscripts

This dissertation was conducted at the *Center for Doctoral Studies in Social Sciences* (CDSS) within the *Graduate School of Economic and Social Sciences* (GESS) at the University of Mannheim. It is based on three manuscripts; one of which has been published, one of which has been invited for revision, and one of which has been submitted for publication. In addition, I refer to three further manuscripts which I contributed to as a co-author; two of them published, one unpublished.

The research projects presented in Manuscripts I to III were supported by Grant GR-4649/4-1 (PA 1925/2-1) from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

The main text summarizes and discusses the three manuscripts with the goal of integrating and broadening our understanding of the factors shaping numerical real-world estimation. Detailed descriptions of experimental procedures and analyses can be found in the appended original manuscripts.

MANUSCRIPT I

Kreis, B. K., Groß, J., & Pachur, T. (2025). Real-world estimation taps into basic numeric abilities. *Psychonomic Bulletin & Review*, *32*(3), 1217-1230. https://doi.org/10.3758/s13423-024-02575-4

MANUSCRIPT II

Kreis, B. K., Notarbartolo, C., Pachur, T. & Groß, J. (2025). *Improving water-footprint estimates and promoting sustainable food choices: A comparison of three simple interventions.* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

MANUSCRIPT III

Kreis, B. K., Hermann, A., Pachur, T. & Groß, J. (2025). *Hindsight bias through knowledge updating: A conceptual replication of Groß et al. (2023).* Manuscript invited for revision at *Collabra: Psychology.*

ADDITIONAL MANUSCRIPTS

In addition to the three manuscripts included in this dissertation, I contributed as a co-author to three further manuscripts on numerical real-world estimation. Although these manuscripts are not part of the dissertation, given that I was not the primary contributor, they are closely related to the overarching research topic and offer valuable insights. As such, I reference them at relevant points throughout this dissertation.

Groß, J., Kreis, B. K., Blank, H., & Pachur, T. (2023). Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General*, *11*(152), 3167–3188. https://doi.org/10.1037/xge0001452

Groß, J., Loose, A. M., & Kreis, B. K. (2024). A simple intervention can improve estimates of sugar content. *Journal of Applied Research in Memory and Cognition*, *13*(2), 282–291. https://doi.org/10.1037/mac0000122

Izydorczyk, D., Kreis, B. K., Kilb, M., & Bröder, A. (2025). *# Knowledge Using social media for improving food-related knowledge: A seeding intervention.* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

1 Introduction

The ability to estimate unknown quantities is a fundamental cognitive skill across species (Nieder, 2019). In the animal kingdom, higher estimation competence is positively associated with various adaptive outcomes, including foraging efficiency, social coordination, predator avoidance, and ultimately, survival and reproductive success (Nieder, 2020). For humans, this competence is particularly crucial, and increasingly so, as daily life in today's Information Age is becoming more and more quantified and data-driven (Boels et al., 2025; Castells, 1997). Every day, we rely on numerical estimation to quickly and effectively assess situations and make decisions. We estimate when to leave the house to catch a bus, how long our phone battery will last, or whether our grocery spending remains within budget. Estimation enables us to approximate such quantitative values when exact figures are unavailable, time is limited, or precise calculations are impractical (Brunswik, 1952, 1955; Gigerenzer & Goldstein, 1996; Goldstein & Hogarth, 1997; Pachur & Bröder, 2013).

As numbers have become increasingly pervasive in daily life, scientific interest in numerical estimation has grown in unison (e.g., Brunswik, 1952, 1955; Gigerenzer & Goldstein, 1996; Goldstein & Hogarth, 1997; Pachur & Bröder, 2013; Tversky & Kahneman, 1974). To gain a deeper understanding of this cognitive skill and its related phenomena, researchers have investigated it across a wide array of disciplines and perspectives, ranging from fundamental to applied research, from exploring the underlying cognitive processes involved in estimation (e.g., Brunswik, 1955; Izydorczyk & Bröder, 2023; Pachur & Bröder, 2013) to its influence on health- and sustainability-related behaviors (e.g., Dallacker et al., 2018; Marghetis et al., 2019), and spanning fields such as cognitive science, psychology, health, and environmental studies. This has resulted in a diverse and extensive body of research on numerical estimation. However, as studies across fields have developed largely in isolation, the literature remains conceptually and methodologically fragmented. This fragmentation poses a significant challenge: when valuable insights remain siloed, the potential for cumulative scientific progress is diminished. As Isaac Newton has famously remarked in a letter to Robert Hooke in 1675, "If I have seen further, it is by standing on the shoulders of giants", a reminder that scientific advancement relies on building upon and synthesizing existing knowledge.

For numerical estimation, this lack of integration leaves us without a clear understanding of the factors that shape this skill, especially in real-world contexts. A key task for research on numerical estimation is, therefore, to foster integration. This dissertation is dedicated to that goal. To gain a more comprehensive understanding of the factors underlying numerical estimation—particularly in real-world contexts—I adopt an integrative perspective and employ a methodological approach that combines experimental methods with investigations using real-world content. This approach offers unique advantages: experimental paradigms provide internal validity and allow for a systematic examination of effects and their underlying mechanisms, whereas real-world stimuli reflect the complexity of everyday environments, thus enhancing ecological validity and practical relevance.

This dissertation is organized into four chapters, with Chapter 1 serving as the introduction. In Chapter 2, I address two mental resources that fundamentally shape real-world estimation. Whereas prior research mainly focused on varying aspects of domain knowledge (e.g., Brown, 2002; Brown & Siegler, 1993; Brunswik, 1955; Gigerenzer & Goldstein, 1996; Griffiths & Tenenbaum, 2006; Hammond & Stewart, 2001; Juslin et al., 2003; Tversky & Kahneman, 1974), I extend this line of work by considering the role of basic numeric abilities. Through integrating insights from the broader field of numerical cognition, I demonstrate that symbolic-number mapping—the ability to accurately map symbolic numbers onto a mental number line—exerts a reliable influence on estimation accuracy (Manuscript I).

In Chapter 3, I examine how estimation is shaped by changes in the underlying resources, specifically by alterations to the knowledge base induced through the provision of information. First, building on prior experimental research, I consider how improving the knowledge that underlies numerical estimation through the targeted provision of information can enhance estimation accuracy. Extending this work, I integrate insights from applied intervention research to compare the effectiveness of different informational interventions and examine whether the acquired knowledge translates into improved decision-making (Manuscript II). Second, I address how the provision of information does not only improve estimates but can also unintentionally distort the recall of estimates. This phenomenon, known as hindsight bias, refers to the tendency to recall previous estimates as more accurate than they actually were after learning the actual value (Blank et al., 2007; Tversky & Kahneman, 1974). Building on a co-authored paper (Groß et al., 2023), I argue that viewing hindsight bias from the perspective of numerical estimation research reveals that

both phenomena of altered estimation, improved estimation and distorted recall, appear to stem from the same source: the updating of knowledge.

In Chapter 4, I synthesize the findings from the individual manuscripts and discuss their theoretical and practical implications. I highlight the value of integrating diverse lines of research, discuss open questions and outline how future work can continue to broaden and integrate our understanding of numerical real-world estimation.

2 Foundations of Numerical Estimation

As outlined in the Introduction, the ability to estimate unknown quantities is a fundamental cognitive skill that we frequently rely on in everyday life. But what are the foundations of this ability? Which mental resources do people rely on when estimating unknown numerical quantities in real-world contexts?

Consider the following example: someone participating in a quiz is asked to indicate the population of Spain. Most people are unlikely to know the exact figure and will therefore need to generate an estimate. Suppose the person responds with 60 million inhabitants. Although this estimate is not entirely accurate, with the actual population being approximately 48 million, it is reasonably close. How might they have arrived at this figure? It is likely that the individual drew on various pieces of information and integrated them into a single numerical estimate (e.g., Brown, 2002; Brown & Siegler, 1993; Brunswik, 1955; Gigerenzer & Goldstein, 1996; Goldstein & Hogarth, 1997; Griffiths & Tenenbaum, 2006; Tversky & Kahneman, 1974). To begin with, they might know that most country populations range from tens to low hundreds of millions. They may also recognize that Spain is a large, but not the most populous European country. Further recalling that Germany, the most populous country of the European Union, has around 80 million inhabitants, they may infer that Spain's population should be somewhat lower than that number. By combining these insights, the person could then arrive at an estimate of 60 million. Now consider a second question about the population of Ethiopia. Using a similar strategy, the individual may reason that Ethiopia is a somewhat well-known country and therefore unlikely to be very small. However, being aware that they lack more specific knowledge, they might estimate its population to be slightly lower than Spain's, say, around 40 million. This, however, now substantially underestimates the actual figure, which is nearly 129 million.

This example illustrates what is arguably the most fundamental mental resource people draw on when estimating: *domain knowledge* (Brown, 2002; Brown & Siegler, 1993). Depending on its accuracy, specificity, and relevance, the knowledge that an individual possesses can significantly shape the precision of the resulting estimates

(e.g., Brown, 2002; Brown & Siegler, 1993; Brunswik, 1955; Gigerenzer & Goldstein, 1996), as the contrast between our hypothetical estimates of Spain and Ethiopia suggests. This observation is not only intuitive but also reflects a longstanding focus in the study of numerical estimation. Although different traditions use varying terminology and conceptual frameworks, they share a common assumption: domain knowledge—be it learned associations (Brunswik, 1955; Hammond, 1955; Hoffmann et al., 2019), retrieved exemplars (Izydorczyk & Bröder, 2021; Juslin et al., 2003), heuristics (Gigerenzer & Goldstein, 1996; Tversky & Kahneman, 1974), or statistical abstractions (Griffiths & Tenenbaum, 2006; Lewandowsky et al., 2009)—is a highly relevant factor in guiding and shaping numerical estimates.

In the following section, I will provide a brief overview of those perspectives. Building on this foundation, I then argue that a sole focus on knowledge may be, however, incomplete. Following this argument, in Manuscript I, I demonstrate that basic numeric abilities should also be considered as a foundational mental resource shaping numerical real-world estimation.

2.1 Domain Knowledge

One of the earliest lines of research to consider the role of what I broadly refer to as domain knowledge in forming numerical estimates is the Brunswikian research tradition (Brehmer & Brehmer, 1988; Brunswik, 1952, 1955; Doherty & Kurz, 1996; Goldstein & Hogarth, 1997; Hammond, 1955; Hammond & Stewart, 2001). This line of research proposes that individuals estimate unknown quantities by drawing on learned abstracted rules of how the to-be-estimated criterion quantities (e.g., country populations) are related to observable cues (e.g., that larger geographical size is correlated with a greater population). Building on these ideas, a substantial body of cue-utilization research has since developed, with an increased focus in recent years on using formal models to explain how cue-criterion knowledge is structured and integrated during estimation (e.g., Einhorn et al., 1979; Juslin et al., 2003; Medin & Schaffer, 1978; Pachur & Bröder, 2013). In this context, two key perspectives have emerged on how cue-criterion knowledge is structured and integrated: rule-based and exemplar-based approaches (e.g., Hintzman, 1986; Izydorczyk & Bröder, 2023; Medin & Schaffer, 1978; von Helversen & Rieskamp, 2009). Rule-based models, closely related to the Brunswikian tradition, emphasize knowledge in the sense of abstract probabilistic rules of systematic associations between cues and criterion

values (e.g., Hoffmann et al., 2019; Izydorczyk & Bröder, 2023; Juslin et al., 2003; von Helversen & Rieskamp, 2009). In contrast, exemplar-based models assume that estimates are guided by memory for specific, previously encountered instances (e.g., Hintzman, 1986; Hoffmann et al., 2013; Juslin et al., 2003; Medin & Schaffer, 1978). Rather than relying on abstract rules, these models posit that individuals retrieve known exemplars (e.g., France, Germany) and estimate the quantity of unknown objects based on their similarity to these known exemplars.

Whereas research on cue-utilization emphasizes how structured knowledge is logically combined to form estimates, the heuristics-and-biases framework offers a complementary perspective. It focuses on how intuitive mental shortcuts or heuristics, in a sense more intuitive and qualitative knowledge, shape numerical judgments (Tversky & Kahneman, 1974). A prominent example of such a heuristic that influences numerical estimation is the availability heuristic. It suggests that people rely on the ease with which examples or relevant information come to mind when estimating unknown quantities, estimating objects that come more easily to mind (i.e., that are more cognitively "available") as larger (Brown & Siegler, 1993; Tversky & Kahneman, 1973). Originally, such heuristics were characterized as problematic, leading to systematic judgment errors or biases (e.g., our quiz-taker underestimating Ethiopias's population relative to Spain simply because more information about Spain was readily accessible in their memory, Tversky & Kahneman, 1973, 1974). Later research, however, has emphasized the efficiency and ecological rationality of using such qualitative knowledge (e.g., Chater & Oaksford, 2000; Gigerenzer & Goldstein, 1996; Pachur & Bröder, 2013). If the highly abstracted information of a heuristic matches the structure of the environment, relying on these strategies can support accurate estimation. For example, being asked about the population size of the Union of Comoros, our quiz-taker might assume that this country they have never heard of likely has a very small population. In this case, the realization of an absence of available knowledge would serve as a valid cue: the population is indeed less than one million.

Another form of knowledge that is highly relevant to numerical estimation is that of the statistical properties of a domain. Research suggests that through repeated exposure to objects of a domain and their associated quantities, people often develop relatively accurate intuitions about a domain's distributional features, such as the typical range, overall form (e.g., kurtosis, skewness), or central tendency of values within that domain (Griffiths & Tenenbaum, 2006; Lewandowsky et al., 2009;

Spencer, 1961). For instance, people who have encountered demographic information in geography classes, travel books, or quiz games may have a general sense that most country populations range from a few million to a few hundred million, with only a few outliers.

A particularly explicit account of how different forms of knowledge inform numerical estimation is offered by the *metrics and mapping framework* by Brown and Siegler (1993). Aimed at integrating insights from prior research and adapting them specifically to real-world estimation contexts, this framework offers a practical synthesis. The framework distinguishes between two complementary forms of knowledge that shape estimation: metric knowledge and mapping knowledge (Brown & Siegler, 1993). *Metric knowledge* refers to an understanding of the statistical structure of a domain, such as its typical range, central tendency, and distribution (e.g., knowing that most country populations fall between a few million and a few hundred million). *Mapping knowledge*, by contrast, refers to knowledge of the relative ordering of objects within a domain. This includes heuristic intuitions, such as those captured by the availability heuristic, as well as knowledge of the relationship between specific exemplars (e.g., knowing that Germany has a larger population than Spain). These two knowledge types are considered conceptually independent (Brown & Siegler, 1993). One can know the approximate range of country populations (metric knowledge) without knowing how two specific countries compare (mapping knowledge), and vice versa. At the same time, they are assumed to jointly guide numerical estimation: metric knowledge provides the numerical response range, whereas mapping knowledge determines where within that range a specific object is placed (Brown, 2002; Brown & Siegler, 1993).

Taken together, although these diverse research strands emphasize different aspects of knowledge, they converge on a key assumption: acquired knowledge is a key resource that guides people's attempts to estimate unknown quantities.

2.2 Basic Numeric Abilities

Kreis, B. K., Groß, J., & Pachur, T. (2025). Real-world estimation taps into basic numeric abilities. *Psychonomic Bulletin & Review*, *32*(3), 1217-1230. https://doi.org/10.3758/s13423-024-02575-4

Although prior research has provided valuable insights into how various forms of knowledge guide numerical estimation, in Manuscript I (Kreis, Groß, & Pachur, 2025), we argue that this focus overlooks a potentially crucial component: basic numeric abilities. In this context, we use the term basic numeric abilities to refer to foundational skills related to understanding and manipulating numbers in an abstract, content-independent manner—that is, an understanding of the numbers themselves. In particular, in Manuscript I, we explore the role of the basic numeric ability *symbolic-number mapping* as a complementary mental resource that may shape numerical real-world estimation.

Symbolic-number mapping refers to the ability to accurately locate numbers on a mental number line and represent their magnitudes proportionally relative to each other (Schneider et al., 2018; Siegler & Opfer, 2003; Thompson & Opfer, 2016). It is typically assessed using the number-line task (Siegler & Opfer, 2003), where participants are asked to place a given number (e.g., 486) onto a blank horizontal line bounded by a defined start and endpoint (e.g., 0 to 1,000; see bottom row in Figure 1). Symbolic-number mapping has been extensively studied in the field of numerical cognition as an important ability underlying performance in various complex numerical tasks (e.g., Peters & Bjalkebring, 2015; Schley & Peters, 2014; Schneider et al., 2018; Thompson & Siegler, 2010). Studies have shown that, for instance, more accurate symbolic-number mapping is associated with better memory for numerical information (Peters & Bjalkebring, 2015; Thompson & Siegler, 2010), proportionally more accurate subjective value assessments of objective magnitudes (Schley & Peters, 2014), and more normatively appropriate decisions in risky choice scenarios (Patalano et al., 2020; Peters & Bjalkebring, 2015).

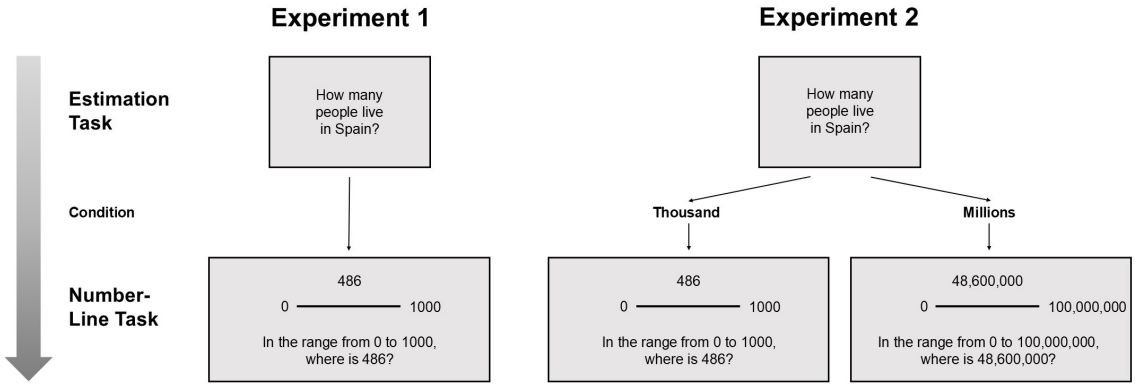
Nonetheless, despite its well-established relevance across various numerical tasks, symbolic-number mapping has not yet been examined in the context of numerical real-world estimation. However, there are good reasons to assume it may be relevant. As Brown (2002) argues, during real-world estimation individuals first generate a plausible metric response range and then locate the target object within that range,

effectively mapping an object to a position along a mental scale before converting this position into a numerical estimate. This procedure thus closely parallels the cognitive operations involved in symbolic-number mapping.

The aim of Manuscript II was therefore to investigate the potential role of symbolic-number mapping in numerical real-world estimation, specifically whether participants' symbolic-number mapping ability was associated with their estimation accuracy. To this end, we conducted two preregistered online experiments. In both experiments, participants first completed an estimation task, followed by a number-line task (for an overview of the procedure of the experiments, see Figure 1).

Figure 1

Procedure and Design of the Experiments in Manuscript I (adapted from Kreis, Groß, & Pachur, 2025)



Note. Thousand = Thousand condition. Millions = Millions condition.

In the estimation task, participants estimated the population sizes of various countries, a domain commonly used in research on numerical real-world estimation (e.g., Brown & Siegler, 1993, 1996; LaVoie et al., 2002). We quantified their estimation accuracy in terms of the order of magnitude error (OME). The OME (see Equation 1 below) is the logarithmic deviation between participants' estimates and the actual values, and is especially suited for the study of often skewed real-world domains, as it reduces the otherwise inflated influence of outliers (e.g., Bröder et al., 2023; Brown, 2002; Brown & Siegler, 1996; Groß et al., 2023).

Formally, we computed the OME for each item i and each participant j as:

$$\text{OME}_{ij} = |\log_{10}\left(\frac{\text{estimate}_{ij}}{\text{actual}_i}\right)|. \quad (1)$$

Hereby, larger OME values indicate greater error, and therefore lower estimation accuracy.

In the number-line task, participants mapped a series of numbers onto a horizontal line bounded by defined start and end values. Symbolic-number mapping accuracy for each participant j was quantified as the median of the relative deviations of the number mapped on the line and its target value for each item k (see Equation 2 below; Schneider et al., 2018):

$$\Delta_{kj} = |\left(\frac{\text{estimate}_{kj} - \text{actual}_k}{\text{actual}_k}\right)|. \quad (2)$$

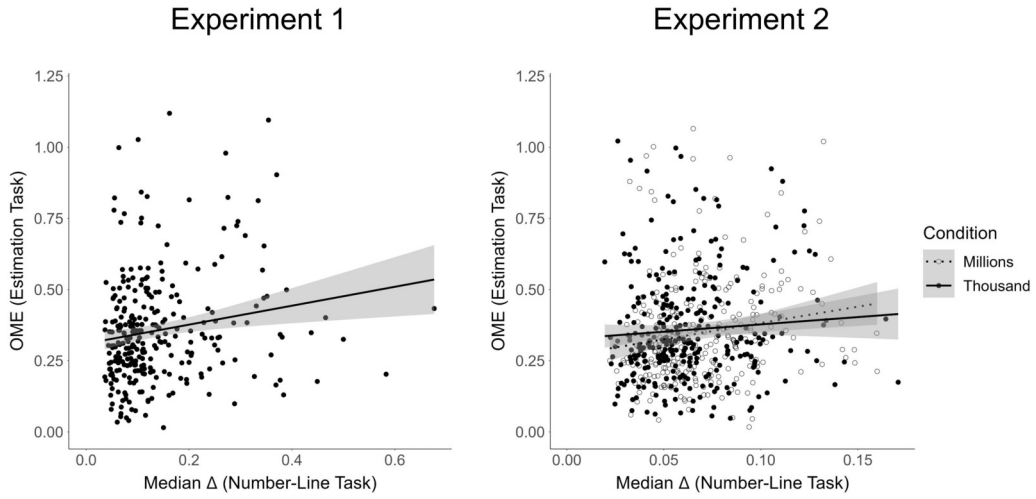
Here, larger median deviations reflect lower symbolic-number mapping accuracy.

In Experiment 1 ($N = 286$), in line with established practices in symbolic-number mapping research, we used a number-line task ranging from 0 to 1,000 (Opfer & Siegler, 2007; Schley & Peters, 2014; Siegler et al., 2009). Using Bayesian hierarchical regression analyses, we demonstrated for the first time that more accurate symbolic-number mapping is associated with more accurate estimates of country populations (see left panel in Figure 2). With Experiment 2 ($N = 592$), we then sought to examine the robustness of these findings and to further investigate the unique role of symbolic-number mapping in numerical real-world estimation. Using the same overall design as in Experiment 1, we first replicated the previously observed association between participants' symbolic-number mapping ability and estimation accuracy. Extending the experimental setup, we further introduced a key modification to the number-line task: Participants either completed the classical version ranging from 0 to 1,000 (as in Experiment 1, *Thousand* condition) or a modified version ranging up to 100,000,000 (*Millions* condition, see Experiment 2 in Figure 1). This allowed us to test whether the numerical mismatch between the estimation task (which involved values in the hundreds of millions) and the number-line task (limited to 1,000 in Experiment 1) affected the observed association. However, results showed that the range of the number-line task did not moderate the relationship between symbolic-number mapping and estimation accuracy (see right panel in Figure 2). This finding suggests that symbolic-number mapping generally reflects

an ability that contributes to estimation accuracy regardless of the specific numerical range used to assess symbolic-number mapping. Finally, with Experiment 2 we further disentangled the influence of symbolic-number mapping from that of domain knowledge. By including a measure of participants' self-reported prior engagement with the topic of country populations as an indicator of acquired domain knowledge, and statistically controlling for this factor, we demonstrated that symbolic-number mapping contributed uniquely to estimation accuracy, beyond the effect of domain knowledge.

Figure 2

Association Between Performance in the Estimation Task and Performance in the Number-Line Task (adapted from Kreis, Groß, & Pachur, 2025)



Note. Each point represents the performance of a participant. For the OME, the median (across items) for each participant is shown. For both OME and median Δ, larger values indicate worse performance. The uncertainty band around the regression line represents the standard error. OME = Order of magnitude error.

This work lays important groundwork for future research into the role of basic numeric abilities, particularly symbolic-number mapping, in guiding numerical real-world estimation. Given that our study focused exclusively on country population estimates, future work should explore whether the strength and nature of this association might vary depending on the examined domain. It is plausible, for example, that in more familiar domains, where people can draw on rich domain knowledge (e.g., sugar content of food; Groß et al., 2024), symbolic-number mapping might be less influential than in unfamiliar domains where such knowledge is

lacking (e.g., water footprint of food; Kreis, Notarbartolo, et al., 2025). Further research is also needed to investigate the specific cognitive mechanisms through which symbolic-number mapping affects estimation performance. Such insights will help refine theoretical models and contribute to a more comprehensive account of the mental resources underpinning numerical estimation.

To conclude, with Manuscript I, I demonstrate that both domain knowledge and basic numeric abilities shape numerical real-world estimation. In doing so, I situate numerical estimation more firmly in the context of its psychological and cognitive foundations, bridging research on numerical cognition and numerical real-world estimation.

3 Malleability of Numerical Estimation

In the previous chapter, I examined the mental resources that fundamentally shape numerical real-world estimation. These resources, however, are not static: they can change through experience and learning (Brown & Siegler, 1993; Fitzsimmons et al., 2023; LaVoie et al., 2002; Opfer & Siegler, 2007; Thompson & Opfer, 2016). Consequently, one would expect that changes in the underlying resources should also lead to corresponding changes in the estimates they shape. For basic numeric abilities, which we first introduced as a novel resource relevant to numerical estimation in Manuscript I, this remains a theoretical proposition that I will return to in the Discussion. For domain knowledge, by contrast, the link between changes in the resource and changes in estimation is already well established, particularly as a consequence of the provision of information (e.g., Brown & Siegler, 1993, 1996; LaVoie et al., 2002).

Consider, for example, our quiz participant who learns after the quiz that Spain has approximately 48 million inhabitants. The next time they encounter a similar estimation task, they are likely to estimate the population of Spain more accurately. Such learning effects are a well-documented consequence of the provision of factual information and are typically attributed to the updating of the knowledge base that underlies estimation (e.g., Brown & Siegler, 1993, 1996, 2001; LaVoie et al., 2002). Targeting and updating knowledge as a key mental resource of estimation thus represents a promising approach to improve real-world estimation. However, although this approach has been thoroughly investigated in cognitive psychology (e.g., Brown & Siegler, 1993; LaVoie et al., 2002; Murray & Brown, 2009), its application to real-world contexts, where better estimates could foster more informed decisions, remains limited. In the first part of this chapter, presenting Manuscript II (Kreis, Notarbartolo, et al., 2025), I seek to bridge this gap through integrating experimental insights with applied research approaches. I delineate the potential of different types of interventions, each targeting different aspects of knowledge, and compare their effectiveness in improving estimation accuracy and adaptive decisions.

In the second part of this chapter, presenting Manuscript III (Kreis, Hermann, et al., 2025) I turn to a different, often unintended consequence of the provision of information: the distortion of memory for prior estimates. Research shows that once people learn the actual value of an estimate, they tend to recall their initial estimate as being closer to the true value than it was, a phenomenon known as hindsight bias (e.g., Blank et al., 2007; Christensen-Szalanski & Willham, 1991; Roese & Vohs, 2012; Tversky & Kahneman, 1974). Returning to our quiz example, the participant might later recall having guessed 55 million for Spain's population, closer to the correct figure of 48 million, despite originally estimating 60 million. Traditionally investigated separately from numerical estimation research, I argue that hindsight bias also reflects numerical estimation being shaped by changes in knowledge. Building on the idea that hindsight bias may result from the same knowledge-updating process underlying learning effects, I present research that unites both areas to systematically test this hypothesis (Groß et al., 2023; Hoffrage et al., 2000).

3.1 Improvement of Estimates

The improvement of numerical real-world estimation through the provision of information has been extensively studied using the *seeding paradigm* developed by Brown and Siegler (1993). In this approach, participants are presented *seed facts*, that is actual values of objects of a domain, to examine how such information affects subsequent estimates. To explain these effects, Brown and Siegler (1993) draw on their metrics and mapping framework, linking changes in estimation accuracy to changes in the underlying knowledge base.

In the seeding paradigm, participants first estimate the quantities of a series of real-world items within a given domain (e.g., country populations). During the subsequent seeding phase, they are presented with the actual values for a subset of items, i.e., seed facts (e.g., "The population of Spain is 48 million"). Finally, participants are asked to estimate both the previously seeded items again and a set of new, unseeded, transfer items. This design allows researchers to distinguish between two types of learning effects and thereby differences in the comprehensiveness of knowledge updating. Improved estimates for seeded items reflect *direct learning effects* which indicate local, item-specific updating. In contrast, improved estimates for unseeded transfer items reflect *transfer learning effects*, which suggest that the newly acquired knowledge has been generalized to novel objects.

A series of studies has demonstrated that seeding the knowledge base improves estimation accuracy through eliciting an updating of the underlying knowledge in various real-world domains, including country populations (Brown & Siegler, 1993; LaVoie et al., 2002), geographic coordinates (Friedman & Brown, 2000), automobile prices (Murray & Brown, 2009), and college tuition fees (Lawson & Bhagat, 2002). The scope of these improvements, however, differs between metric and mapping knowledge. For metric knowledge—typically quantified as the deviation between estimated and actual values, often via the OME (see Equation 1)—research consistently shows both direct and transfer learning effects (e.g., Brown & Siegler, 1993; LaVoie et al., 2002; Murray & Brown, 2009; Wohldmann & Healy, 2020). These improvements further tend to persist over extended periods of time (Brown & Siegler, 1996; LaVoie et al., 2002). This stability suggests that the effects are based on a comprehensive and lasting recalibration of participants' knowledge of the domain's statistical structure (Brown & Siegler, 1993, 1996).

In contrast, early studies suggested that improvements in mapping knowledge—typically assessed via the rank-order correlation between estimated and actual values (Brown & Siegler, 1993)—are more limited. Typically, only rather short lived direct learning effects were observed (e.g., Brown & Siegler, 1993; LaVoie et al., 2002; Wohldmann & Healy, 2020). This was interpreted as evidence that participants struggle to extract generalizable patterns from seed facts that would support a transferable updating of mapping assumptions (Brown, 2002; Brown & Siegler, 1993). However, more recent research suggests that under certain conditions participants can extract such transferable mapping generalizations from seeding. This seems especially likely in domains with distinctive substructures of objects that make relational patterns easier to infer (Bröder et al., 2023; Murray & Brown, 2009). For example, participants have been shown to infer generalized ordering patterns in domains like car prices (Murray & Brown, 2009) or foods' carbon footprint (Bröder et al., 2023), where categories like a car being an SUV or food products being animal-based reliably correspond to higher or lower target values.

Given its consistent effects and its conceptual simplicity, the seeding paradigm has been argued to hold strong promise as a practical tool for improving estimation in real-world contexts, especially in domains where informed estimation plays a key role in everyday decision-making (e.g., Bröder et al., 2023; Brown, 2002; Brown & Siegler, 1993; Groß et al., 2024; Wohldmann, 2015). One of the few studies that has investigated the effectiveness of seeding in such a societally relevant domain is

a recent study I co-authored, focusing on sugar content estimation in food products (Groß et al., 2024). Given the well-documented health risks of excessive sugar consumption (Dallacker et al., 2018; Gupta et al., 2018; Winzer et al., 2021; World Health Organization, 2015), the ability to accurately estimate sugar content is a key component of informed and health-conscious dietary decision-making. In this study, we pursued two main objectives. First, we aimed to capture the current state of people's sugar knowledge by assessing estimation accuracy in terms of both metric and mapping knowledge. This allowed us to obtain a broad overview of participants' estimation performance and to identify potential knowledge gaps. Second, we examined the effectiveness of seeding in this health-relevant context. To do so, we implemented the seeding paradigm, which included an initial estimation phase, a seeding phase, and a post-seeding estimation phase. In the seeding phase, participants were randomly assigned to one out of three experimental groups. Beyond a control group that received no relevant information and the classic seeding condition, where participants were presented with the actual sugar content of selected food items, we also introduced a visual seeding condition. In this condition, seed values were additionally displayed using sugar cubes, providing a tangible representation of the quantities. This extension was designed to test whether visualizing the seed information could enhance the effects of seeding by making abstract quantities more concrete and accessible (Schubbe et al., 2020).

Results revealed that, in their initial estimates, participants exhibited acceptable mapping knowledge, but substantial misperceptions of key statistical properties of the domain, reflected in, among others, a considerable overestimation of the sugar content of food items. Importantly, the classic seeding intervention significantly improved estimation accuracy. As expected, metric knowledge improved for seeded and transfer items, reflected in direct and transfer learning effects. Mapping knowledge also improved for the seeded items but did not generalize to transfer items. This limited generalization may be due to participants' already well-developed prior knowledge, which reduced the room for broader updating. Notably, the additional visual sugar-cube representation did not produce greater improvements than those achieved through seeding of actual values alone.

These findings highlight how viewing estimation as being shaped by metric and mapping knowledge, as well as changes in those knowledge aspects, can meaningfully inform applied health and sustainability research. By providing a means to identify distinct knowledge gaps and offering targeted ways to improve estimation,

the seeding paradigm thus provides a valuable basis for designing and evaluating informational interventions. However, despite the effectiveness and strong conceptual relevance of this approach, its use in applied domains remains rare. Beyond our investigation of sugar estimation, only few studies have explored its potential in other societally relevant real-world contexts (Bröder et al., 2023; Wohldmann, 2015; Wohldmann & Healy, 2020). In Manuscript II, we therefore aimed to bridge this gap by aligning the metrics and mapping framework and the seeding paradigm more closely with applied research and practice.

3.2 Targeting Real-World Improvement

Kreis, B. K., Notarbartolo, C., Pachur, T. & Groß, J. (2025). *Improving Water-Footprint Estimates and Promoting Sustainable Food Choices: A Comparison of Three Simple Interventions.* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

Reducing global water consumption is one of today's most pressing environmental challenges, with food production representing a major driver of freshwater use (Food and Agriculture Organization of the United Nations, 2015, n.d.; United Nations, 2023). Shifting toward less water-intensive diets offers a promising strategy for more sustainable water use (Bajželj et al., 2014; Ercin & Hoekstra, 2014; Springmann et al., 2016). To enable such shifts, individuals need to be able to accurately estimate the water footprint of food, that is, the amount of water used in its production (Hoekstra, 2017; Hoekstra et al., 2012). Unfortunately, although there are indications that consumers lack accurate knowledge in this domain (e.g., Attari, 2014; García-González et al., 2020), empirical research on public knowledge of food-related water use is limited.

To address this gap, our first research aim in Manuscript II was to assess individuals' levels of knowledge regarding the water footprint of food. Specifically, we focused on their metric and mapping knowledge (Brown & Siegler, 1993). Both are fundamental for accurate estimation and, ultimately, for informed decision-making.

To promote sustainable dietary decisions, it is crucial to identify effective ways to improve this knowledge. As previously discussed, the *seeding* intervention (Brown & Siegler, 1993) presents a promising, theory-based approach to do so. Based on prior

findings (e.g., Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024; LaVoie et al., 2002; Wohldmann & Healy, 2020), we expected that exposing participants to the actual water footprints of a representative set of food products (e.g., "The water footprint of 1 kg of tomatoes is 110 liters") would enhance both metric and mapping knowledge, for seeded as well as unseeded transfer items. We anticipated that such a transfer of mapping knowledge is plausible, as the food domain has a clear relational structure: food groups such as vegetables and meats differ reliably in their water footprint. Combined with individuals' likely limited prior exposure to this topic, participants might be able to infer such previously unknown relational regularities that enable broader generalization.

To contextualize and evaluate the usefulness of seeding, we compared it to two informational interventions based in applied sustainability research: a rule-based and a label-based approach (e.g., Camilleri et al., 2019; Egnell et al., 2020; Marghetis et al., 2019; Sonnenberg et al., 2013). With the *rule* intervention, we provided participants with a relational summary of food groups' differences in water footprint ("Fruits and vegetables require little water, while animal products, beans, nuts, and seeds consume more"). Whereas it conveys no numerical information, and thus should not affect metric knowledge, the generalized relational information should improve mapping knowledge and support transfer to a broad range of food products (e.g., Brown, 2002; Brown & Siegler, 1993; Marghetis et al., 2019). With the *label* intervention, we offered simplified, item-level rankings (e.g., "A banana has a water-intensity score of 2 out of 5"). Like the rule intervention, the label intervention does not convey metric information but may improve mapping knowledge for labeled items, and possibly support transfer by revealing underlying relational patterns.

With these three informational interventions, each targeting different aspects of knowledge, the second aim of Manuscript II was to assess how effectively they improved participants' metric and mapping knowledge. A key focus was hereby whether these improvements transferred to new items, a crucial condition for supporting real-world decisions across diverse foods.

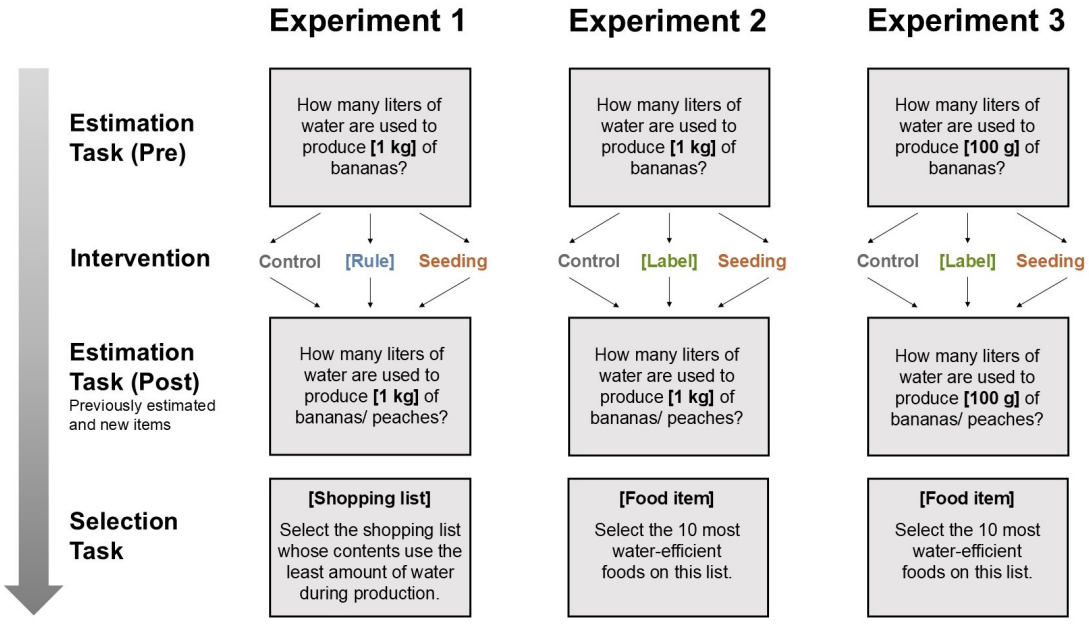
The third aim of Manuscript II was then to test whether such knowledge gains would translate into more sustainable decisions, specifically, selecting the most sustainable option among different food products. This step was essential because, although improving knowledge of the water footprint of food products is a crucial step, its practical relevance depends on whether it ultimately translates into more sustainable decision-making. Theoretically, seeding seemed particularly promising

for supporting sustainable decision-making, as it has the potential to improve both metric and mapping knowledge. However, it has not yet been studied in the context of real-world decision-making. Whether it can support more sustainable food choices thus remains an important open question. Rule- and label-based interventions in contrast, have been shown to support relational understanding and support sustainable or healthy decisions in applied contexts (e.g., Egnell et al., 2020; Lohmann et al., 2022; Marghetis et al., 2019; Sonnenberg et al., 2013).

To address our three research questions—(1) the initial level of participants' metric and mapping knowledge, (2) the effectiveness of three interventions in improving this knowledge, and (3) whether improved knowledge supports sustainable decisions—we conducted three preregistered online experiments (Experiment 1: $N = 116$; Experiment 2: $N = 136$; Experiment 3: $N = 125$). All experiments followed the same general structure (see Figure 3 for an overview).

Figure 3

Procedure and Design of the Experiments in Manuscript II (Kreis, Notarbartolo, et al., 2025)



Note. Marked in brackets are aspects that differ between experiments. Estimation Task (Pre) = Pre-intervention estimation task. Intervention = Intervention phase. Estimation Task (Post) = Post-intervention estimation task. Control = Control group. Rule = Rule intervention. Label = Label intervention. Seeding = Seeding intervention. Shopping list = Shopping-list selection task. Food item = Food-item selection task.

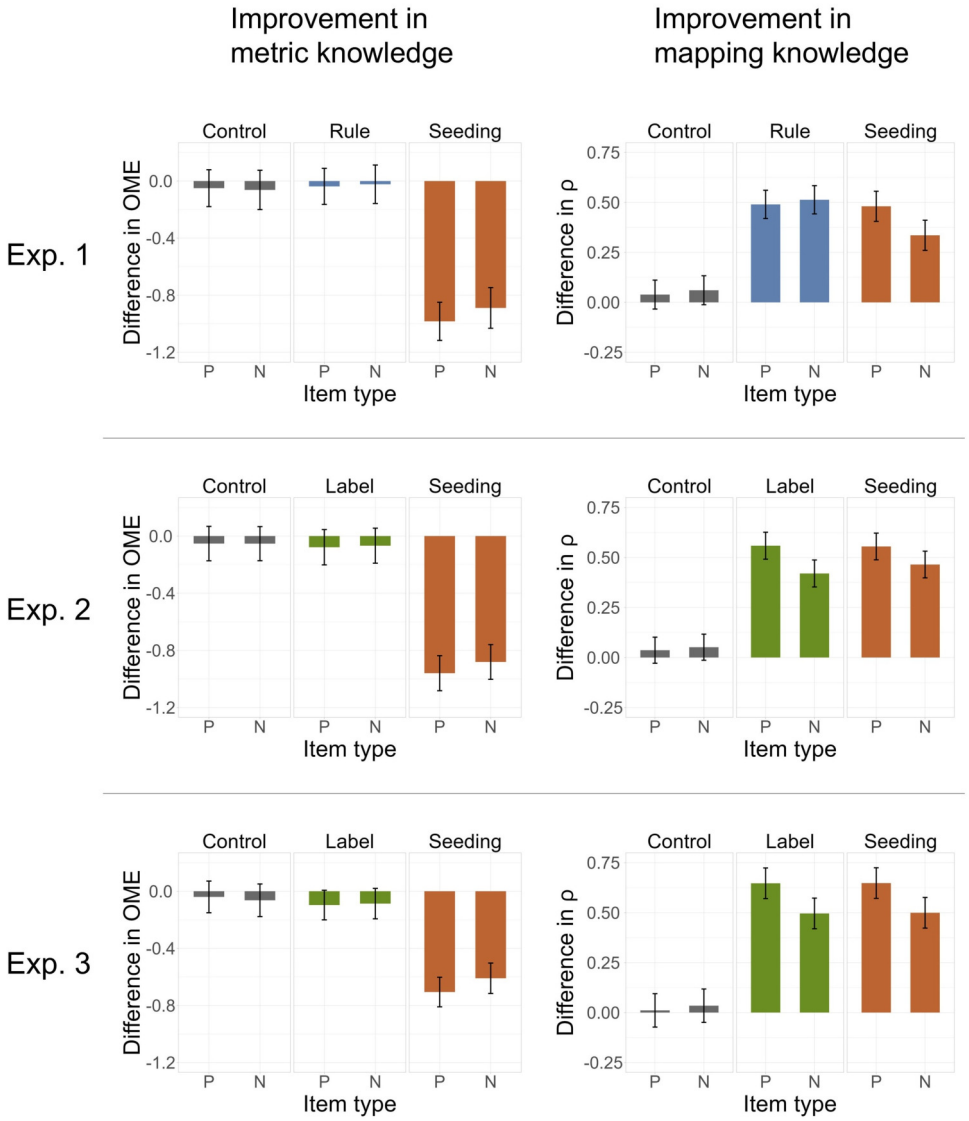
In a pre-intervention estimation task, participants estimated the water footprint of 25 food items to assess their initial levels of metric and mapping knowledge. Next, they were randomly assigned to an intervention group or a control group that read an unrelated text. Besides the control group that read an unrelated text, each experiment included a seeding group and one additional intervention group, either a rule or a label group. In the seeding group, participants were shown the actual water footprint values for the 25 previously estimated items. In the rule group, they received the previously introduced general rule comparing food groups by their typical relative water use. The label group was presented a label for each of the 25 previously estimated items, ranking each item in terms of water intensity (1 = least to 5 = most water intensive). A post-intervention estimation task followed, where participants estimated the 25 previously estimated (and seeded/labeled) items and 25 new ones, to assess both direct and transfer learning effects. Finally, participants completed a selection task to evaluate the effects on sustainable decision-making.

In Experiment 1, we compared the seeding and rule interventions and investigated sustainable decision-making via a shopping-list selection task where participants had to select the least water-intensive list out of four lists, each containing 10 food items. Experiment 2 compared the seeding and label interventions, examining selection performance using a food-item selection task requiring selection of the 10 least water-intensive items from a set of 20. Experiment 3 replicated Experiment 2 but changed the estimation unit from water footprint for 1 kg of food product to 100 g, testing whether such a possibly more relatable portion size might improve estimation accuracy. Similar to prior seeding studies, we quantified metric knowledge as the OME defined in Equation 1 (e.g., Bröder et al., 2023; Brown & Siegler, 1996; LaVoie et al., 2002). Mapping knowledge was quantified as the rank-order correlation ρ between participants' estimates and actual values.

We found that participants initially lacked both metric and mapping knowledge. They substantially underestimated food items' water footprints, for both estimated quantities of food product (1 kg and 100 g) They further showed low alignment with actual rankings, posing a clear barrier to sustainable dietary decision-making. Encouragingly, all interventions effectively improved the targeted knowledge aspects (see Figure 4 for an overview of the change in metric and mapping knowledge).

Figure 4

Effects of the Interventions on Metric and Mapping Knowledge as Manifest in the Estimation Tasks (Kreis, Notarbartolo, et al., 2025)



Note. Shown are contrasts between conditional predictions from the corresponding regression model. Specifically, each bar represents the predicted difference in OME/ρ between the pre-intervention estimation task and the post-intervention estimation task. Contrasts are computed separately for previously estimated (P) and new (N) items. Error bars indicate 95% credible intervals. Exp. 1 = Experiment 1, Exp. 2 = Experiment 2, Exp. 3 = Experiment 3. OME = Order of magnitude error. ρ = rank-order correlation. Control = Control group. Rule = Rule intervention. Label = Label intervention. Seeding = Seeding intervention.

Using Bayesian hierarchical regression analyses, our results showed that the seeding intervention produced the strongest and most comprehensive effects, enhancing estimation accuracy in terms of both both metric and mapping knowledge for previously estimated and new items. The observation that participants indeed showed transfer of mapping knowledge through seeding suggests that they successfully extracted broader relational insights into the underlying structure of food groups and their associated higher or lower water footprints. As expected, the rule and label interventions did not improve metric knowledge, but led to robust gains in mapping knowledge, including previously estimated and new items, also indicating effective transfer.

Regarding participants' performance in the selection tasks, which we analyzed using Bayesian non-hierarchical regression analyses, all interventions led to comparable improvements. This finding demonstrates not only the specific utility of seeding for supporting sustainable decisions, but also more generally underlines the practical utility of improving the knowledge that underlies estimation.

With our experiments, we clearly demonstrate that informational interventions are an effective approach for correcting misconceptions about the water footprint of food and, in turn, supporting more sustainable decisions. Especially the seeding intervention, rooted in cognitive theory, proved to be particularly effective, underscoring both its theoretical value and practical potential. Promisingly, in a recent co-authored study (Izydorczyk et al., 2025) we demonstrated that seeding can be scaled as a public intervention: presenting participants with two seed facts per day on Instagram over a duration of 15 days led to improvements in both metric and mapping knowledge. However, no mapping transfer was found, potentially due to the lower intensity of exposure. In such cases, summarized relational information, such as rules or labels, might complement seeding to further support transfer.

In sum, with Manuscript II, I demonstrate that through targeting the domain knowledge that underlies numerical real-world estimation, simple interventions can effectively improve numerical estimation in societally relevant real-world contexts and support more informed decision-making. Linking theoretical insights on how different aspects of domain knowledge shape estimation with questions of practical applicability, this work thereby bridges cognitive theory and applied sustainability research.

3.3 Biased Recall of Estimates

As briefly introduced at the beginning of this chapter, presenting actual values—as purposefully done in the seeding paradigm—can improve estimation accuracy but also systematically distort memory of one's prior estimates. This phenomenon, known as *hindsight bias*, is robust and widespread (e.g., Christensen-Szalanski & Willham, 1991; Fischhoff, 1975; Guilbault et al., 2004): it has been observed across a wide range of judgment contexts, including event outcome assessments (Fischhoff, 1975), truth judgments (Campbell & Tesser, 1983), judgments of learning (Zimdahl & Undorf, 2021), and numerical estimation (e.g., Bayen et al., 2006; Bernstein et al., 2011; Erdfelder & Buchner, 1998; Groß & Bayen, 2015). In the context of numerical estimation, hindsight bias is typically studied using the so-called memory paradigm (Pohl, 2007). In this paradigm, participants first provide estimates (original judgments; OJ) for a set of items, are then shown the actual values, and later asked to recall their initial estimates (recall of original judgments; ROJ). These recalled judgments are usually shifted toward the actual values, reflecting hindsight bias. This effect has been demonstrated across a broad range of numerical domains, including historical dates, river lengths, building heights, and city populations (e.g., Bayen et al., 2006; Bernstein et al., 2011; Erdfelder & Buchner, 1998; Groß & Bayen, 2015).

Over the past decades, numerous theoretical explanations on the processes driving hindsight bias have been discussed, yet the mechanisms remain only partially understood (see Blank et al., 2007; Hawkins & Hastie, 1990; Hoffrage et al., 2000; Roese & Vohs, 2012, for reviews). Generally speaking, the primary source of hindsight bias has been argued to lie in a biased reconstruction process (Dehn & Erdfelder, 1998; Erdfelder & Buchner, 1998). Whereas some degree of distortion may stem from interference with memory retrieval (i.e., recollection bias), the main driver of the effect seems to be a reconstruction bias (Dehn & Erdfelder, 1998; Erdfelder & Buchner, 1998; Groß & Pachur, 2019). That is, when individuals cannot retrieve their OJ they reconstruct it, with that reconstruction being biased by the actual values.

Several theories have been proposed to explain the mechanisms through which this reconstruction becomes biased. One prominent account is the *anchoring-and-adjustment* perspective (Hawkins & Hastie, 1990; Tversky & Kahneman, 1974). This account argues that during reconstruction the actual value serves as an anchor,

from which people adjust away from insufficiently, resulting in a biased ROJ (e.g., Hawkins & Hastie, 1990; Pohl, 1998; Pohl et al., 2003; Tversky & Kahneman, 1974; Wilson et al., 2021). From this perspective, hindsight bias is attributed to cognitive limitations (Fischhoff, 1975; Hawkins & Hastie, 1990). An alternative account, however, suggests that hindsight bias is actually a by-product of an adaptive mechanism: knowledge updating (Hawkins & Hastie, 1990; Hoffrage et al., 2000). According to the *updating-and-rejudgment* account (Hoffrage et al., 2000), learning the actual outcome prompts individuals to revise the knowledge base that supported their original judgment. When participants later try to reconstruct their OJ, they are assumed to repeat the judgment process, now based on the updated knowledge, which leads to a biased ROJ.

Support for this updating perspective is still limited but promising. In studies by Hoffrage et al. (2000) and Nestler et al. (2012), participants' ROJs showed not only a typical hindsight bias (i.e., a shift toward the actual values) but also more valid cue utilization, following feedback on their OJs. For instance, in Nestler et al. (2012), participants judged personality traits of people based on photographs. After receiving self-assessment feedback from the judged targets, participants' ROJs not only moved closer to the actual values but also more strongly reflected valid visual cues (e.g., neatness indicating conscientiousness). These results suggest that as a result of the feedback, participants showed more accurate cue-utilization for the ROJs, which would be consistent with the idea that the ROJs were based on more accurate, updated knowledge.

Yet, key questions remain. The updating-and-rejudgment account remains conceptually vague concerning the nature of updating and its specific consequences. Is the observed updating limited to items for which actual values were received, or does it generalize to other items (Groß et al., 2023)? And does such updating only bias the recall of judgments, or has it also positive downstream consequences, possibly improving estimation performance (Groß et al., 2023)? To move beyond the conceptual vagueness of earlier work, in a co-authored study with Groß, Blank, and Pachur (2023), we recently developed a more specific framework of knowledge updating in hindsight bias. Through viewing hindsight bias from the perspective of being a phenomenon of numerical estimation that is shaped by changes in the underlying knowledge, we were able to derive specific, testable predictions and thereby conduct a more systematic evaluation of the updating-and-rejudgment account of hindsight bias.

In the next section, I will describe this work (Groß et al., 2023) in detail, highlighting how our integrative approach advances the understanding of the role of knowledge updating in hindsight bias. Following that, I will present Manuscript III, which replicates these findings in a new real-world domain, providing crucial evidence on their robustness and generalizability.

3.4 Understanding Bias Through Improvement

Notably, despite being a phenomenon of numerical real-world estimation, hindsight bias in numerical estimation has primarily been examined from a judgment and memory perspective, largely disconnected from other literature on numerical real-world estimation. However, connecting hindsight bias research with this line of research, specifically, the metrics and mapping framework and findings from seeding research (Brown & Siegler, 1993, 1996), offers fresh insights into the mechanisms driving the bias (Groß et al., 2023).

To recall, seeding research has shown that presenting actual values (i.e. seed facts) prompts reliable updating of domain knowledge, especially regarding a recalibration of metric knowledge, reflected in improved estimates for both seeded items (i.e., direct learning) and unseeded, transfer items (i.e., transfer learning) (Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024). Building on these insights, we proposed that the actual values presented in hindsight bias paradigms function as seed facts (Groß et al., 2023). The provided actual values should thus not only induce hindsight bias but also promote broader knowledge updating, resulting in hindsight bias and learning effects as co-occurring effects of numerical estimation being shaped by an updating of domain knowledge. To empirically test this assumption, we developed an Integrated Hindsight-Bias-and-Seeding Paradigm that builds on the procedural and conceptual parallels between hindsight and seeding paradigms and allows for the simultaneous measurement of both phenomena. In this paradigm, participants first provide original judgments (OJ) of real-world quantities (e.g., country populations). In a subsequent phase they then receive the actual values (i.e., seed facts) for some or all items, paralleling the feedback phase in both paradigms. Afterward, participants recall their original estimates (ROJ), enabling measurement of hindsight bias. Finally, they complete a second estimation task involving previously estimated, seeded items (OJ again) and new items (new judgment, NJ), allowing assessment of direct and transfer learning, respectively.

In a first experiment, participants completed this paradigm for country population estimates. Results confirmed that hindsight bias co-occurred with both direct and transfer learning effects. These findings support the hypothesis that actual values in hindsight paradigms function as seed facts, prompting knowledge updating. The presence of transfer learning effects further suggests that knowledge was broadly updated. However, although these results demonstrate a co-occurrence, they do not conclusively confirm that hindsight bias and learning share the same underlying process; actual values may have triggered both independently.

To test more directly whether it is truly knowledge updating that drives hindsight, we derived two further predictions which we tested in a second experiment (Groß et al., 2023). First, if hindsight bias results from knowledge updating, any information triggering such updating should also induce hindsight bias. We tested this by providing participants with actual values for different objects from the same domain (*domain information*) rather than for the originally estimated items. As prior seeding research shows that domain information prompts broad knowledge updating, reflected in transfer learning (e.g., Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024), if hindsight bias is indeed a product of this updating process, it should also emerge under these conditions alongside transfer learning. Second, and conversely, if knowledge updating is necessary for hindsight bias, irrelevant numerical information (*irrelevant information*) should not elicit hindsight bias. To test this, participants received the actual values for the OJ items, but these were framed as irrelevant information from a different domain (e.g., movie budgets). These values, unrelated to the estimation domain, cannot be used to update domain knowledge, and should therefore produce neither transfer learning nor hindsight bias.

Again measuring hindsight bias and transfer learning using the Integrated Hindsight-Bias-and-Seeding Paradigm with country population estimates, the results confirmed our hypotheses. In the domain-information condition, both transfer learning and hindsight bias were observed. In the irrelevant-information condition, neither hindsight bias nor learning effects were found. These findings are compelling. They provide the first direct evidence that hindsight bias can be triggered not only by providing the actual values of the specific items that were previously estimated, but also by receiving domain-relevant information that elicits broader knowledge updating. Moreover, the absence of effects in the irrelevant-information condition suggests that the observed hindsight bias emerged not due to superficial anchoring to the presented numerical value but reflects genuine knowledge updating.

Kreis, B. K., Hermann, A., Pachur, T. & Groß, J. (2025). *Hindsight Bias Through Knowledge Updating: A Conceptual Replication of Groß et al. (2023).* Manuscript invited for revision at *Collabra: Psychology*

To further solidify and broaden our insights into hindsight bias as a by-product of knowledge updating, Manuscript III (Kreis, Hermann, et al., 2025) presents a conceptual replication of Experiment 2 from Groß et al. (2023). Although closely mirroring the structure and logic of the original design, this replication introduces a critical change in content. Instead of estimating country populations, participants estimated the sugar content of food items, a domain that is likely more familiar and accessible in everyday life.

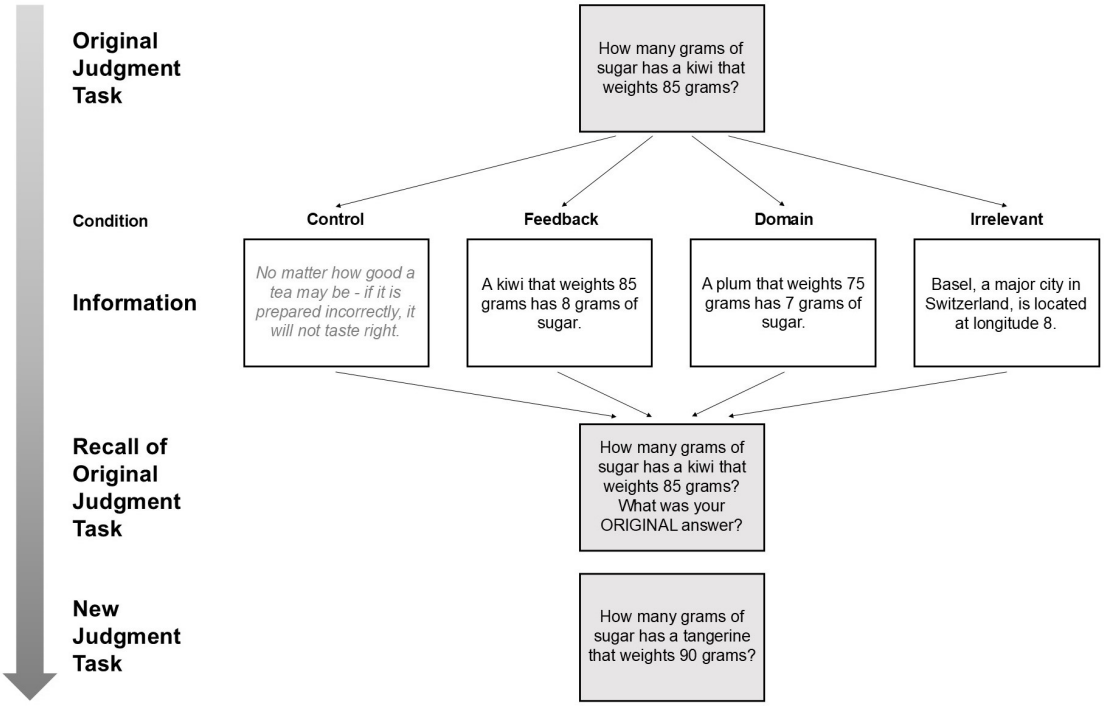
This replication served two main aims. First, it allowed us to assess the robustness of the original findings. Given their theoretical significance for deepening our understanding of the phenomenon hindsight bias, replication is essential for establishing a solid empirical foundation for theory development and for addressing broader concerns about replicability in psychological science (e.g., Earp & Trafimow, 2015; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Second, the study tested whether the link between knowledge updating and hindsight bias generalizes to a different, more familiar estimation domain. This second aspect matters because the estimation domain investigated in Groß et al. (2023), country populations, is rather seldomly encountered in everyday life. Outside of educational settings, people rarely engage with population data. Familiarity with a domain, however, can shape memory, judgment, and learning, and thus likely also knowledge updating and hindsight bias (e.g., Bellana et al., 2021; Christensen-Szalanski & Willham, 1991; Hertwig et al., 2003). Yet, it remains unclear whether both are equally affected by familiarity, or whether familiarity might alter the relationship between them.

To address these aspects of robustness and generalizability, we conducted a preregistered online experiment ($N = 300$) that closely followed the design of Experiment 2 in Groß et al. (2023). We modified only the stimulus domain, now using sugar content in food products as the target quantities, a domain with which people are likely more familiar with due to regular encounters through nutrition labels, packaging, and health messaging.

In the experiment, we employed the same Integrated Hindsight-Bias-and-Seeding paradigm (see Figure 5 for an overview) as in Experiment 2 in Groß et al. (2023). Participants first completed an initial estimation task (OJ task), where they estimated the sugar content of a variety of food items. They were then randomly assigned to one of four experimental conditions that were identical in structure and logic to those used in Experiment 2 in Groß et al. (2023), each involving different types of informational input. In the *feedback* condition, participants received the actual sugar content of the same food items they had just estimated. In the *domain-information* condition, they were shown actual sugar values for a different set of food items. In the *irrelevant-information* condition, participants saw the same numerical values as the feedback group, but these were relabeled as representing longitudes of European cities. Finally, in the *control* condition, participants read an unrelated text. Subsequently, all participants were asked to recall their original estimates (ROJ task), followed by a final estimation task involving new items (NJ task).

Figure 5

Procedure and Design of the Experiment in Manuscript III



Note. Control = Control group, Feedback = Feedback group, Domain = Domain-information group, Irrelevant = Irrelevant-information group.

To quantify estimation accuracy in each estimation task, OJ, ROJ, NJ, we computed the absolute deviation $|\Delta|$ between each estimate i and the corresponding actual value for each participant j:

$$|\Delta_{ij}| = |\text{estimate}_{ij} - \text{actual}_i| \tag{3}$$

Hereby, a smaller $|\Delta|$ indicates a higher estimation accuracy.

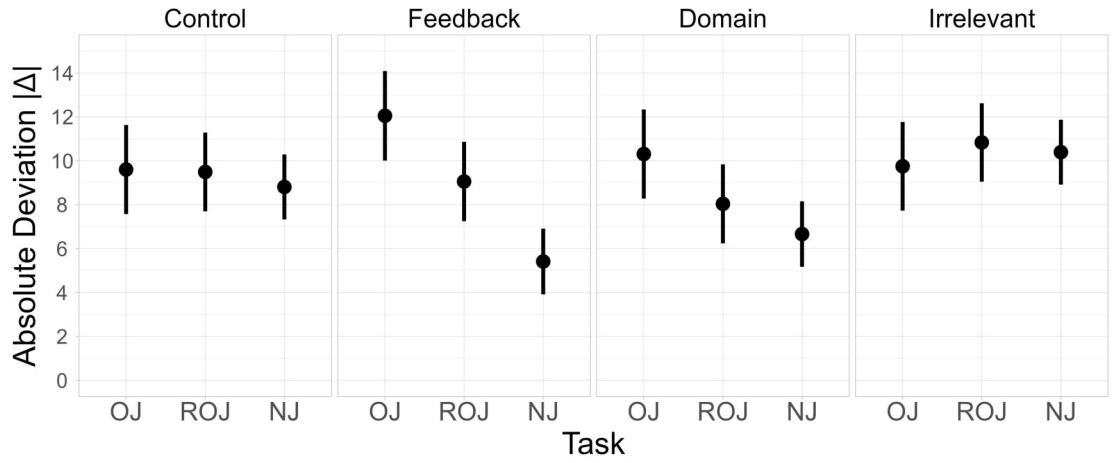
Unlike in Experiment 2 of Groß et al. (2023), in which we used the logarithmic deviation OME as the quantification of estimation accuracy, for this experiment we preregistered $|\Delta|$ due to the less skewed distribution of sugar content values and the frequent presence of zeros in both estimates and actual values, which are incompatible with log-transformation.

To analyze our data, as a first step, we confirmed that participants were indeed more familiar with the domain of sugar content than with that of country populations. Participants' higher familiarity with the sugar content of food products compared to country populations was apparent in several indicators, such as higher initial rank-order correlation, a greater proportion of perfectly accurate estimates in the OJ phase, quicker responses, as well as higher self-reported prior engagement with the domain of foods' sugar content compared to the country population domain. Despite this increased familiarity, the results closely replicated the key findings from Experiment 2 in Groß et al. (2023). Applying Bayesian hierarchical regression models, our analyses revealed the predicted differential pattern. Both the feedback and domain-information conditions produced hindsight bias and transfer learning effects, as reflected in reduced $|\Delta|$ scores from the OJ to the ROJ/NJ tasks (see Figure 6). In contrast, no such effects emerged in the irrelevant-information or control conditions, where estimation accuracy remained stable. This pattern indicates that actual values only triggered hindsight bias when they were usable to update participants' knowledge. Together, these findings demonstrate that the link between knowledge updating and hindsight bias generalizes to the more familiar domain of food products' sugar content, underscoring the robustness of this mechanism across domains with varying familiarity.

Beyond enabling a simultaneous investigation of hindsight bias and transfer learning effects in the same paradigm, our integrated paradigm further allows for a direct comparison of their magnitudes. This opens up valuable opportunities for refining our understanding of the specific mechanisms underlying those phenomena.

Figure 6

Changes in Estimation Accuracy between Estimation Tasks (Kreis, Hermann, et al., 2025)



Note. Shown are the conditional predictions based on the mixed-effects model (estimated means and 95% credible intervals). OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment. Control = Control group, Feedback = Feedback group, Domain = Domain-information group, Irrelevant = Irrelevant-information group.

When comparing the magnitudes of hindsight bias and transfer learning effects, an interesting pattern emerges: in both Groß et al. (2023) and Manuscript III, hindsight bias was consistently smaller than the corresponding transfer learning effects. That is, ROJs were further from the actual values than NJs. However, if hindsight bias solely reflects a re-judgment based on updated knowledge, one would expect hindsight bias to not be smaller than the corresponding learning effects, since both are assumed to rely on the same updated knowledge base. The fact that this is not the case suggests that additional processes beyond knowledge updating contribute to hindsight bias. This assumption that additional processes underlie ROJs does seem plausible. After all, recalling a previous estimate reflects a different demand than simply providing a new estimate. In the ROJ task, participants may therefore engage in additional processes, such as metacognitive reflection on their prior accuracy, alongside the use of updated knowledge.

Our integrative perspective on hindsight bias as a phenomenon of numerical estimation being shaped by changes in knowledge also provides us with additional viewpoints that further deepen our understanding of the specific influence of knowledge updating on the bias. In our main analyses we focused on changes in estimation accuracy in terms of the difference between estimates and actual values, which in

the seeding literature is understood as reflecting metric knowledge (e.g., Brown & Siegler, 1996; Groß et al., 2024; LaVoie et al., 2002). However, it is also informative to consider rank-order accuracy, which captures mapping knowledge (e.g., Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024). Recall that prior seeding research often showed that rank-order accuracy improves only for seeded items, without generalizing to transfer items (e.g., Brown & Siegler, 1993; Groß et al., 2024; LaVoie et al., 2002). If hindsight bias depends on knowledge updating, we would expect a change in participants' ROJ rank-order accuracy only for items that were directly seeded. That means we would expect hindsight bias only in conditions where participants received actual values for the specific items they later recall (feedback condition), but not when such direct information was absent (domain-information condition). This prediction was confirmed in Experiment 2 of Groß et al. (2023). No transfer learning effects were observed, neither in the feedback nor the domain-information condition. Regarding hindsight bias, only in the feedback condition an effect emerged, that is, where participants were presented the actual values for the specific ROJ items. In the domain-information condition, no effect emerged. In Manuscript III, we replicated the hindsight bias in rank-order accuracy in the feedback condition. However, we also observed transfer learning effects in both feedback and domain-information conditions. This was unexpected, as in our prior study on seeding effects in the estimation of foods' sugar content we did not find transfer learning effects for mapping knowledge (Groß et al., 2024). However, consistent with the updating-and-rejudgment account (Hoffrage et al., 2000), hindsight bias in mapping knowledge then also emerged in both conditions. These findings strengthen the idea that hindsight bias is rooted in knowledge updating. The fact that hindsight bias in mapping knowledge only appeared when knowledge was demonstrably updated in terms of learning effects, provides further compelling evidence for the mechanisms proposed by the knowledge-updating account.

To conclude, with Manuscript III, I demonstrate evidence that hindsight bias reflects the same processes of knowledge updating that underlie improved estimates. In showing its generalizability beyond the original stimulus domain, I address a potential boundary condition of this finding, documenting its robustness. With this work, I illustrate how integrating research on previously separately studied phenomena, hindsight bias and seeding effects, can deepen our understanding of these phenomena specifically and of the shaping influence of domain knowledge on numerical estimation in general.

4 Discussion

In this dissertation, I set out to advance our understanding of the factors shaping numerical real-world estimation through integrating and extending existing research traditions. In Manuscript I (Kreis, Groß, & Pachur, 2025), I extended the current understanding of the mental resources that constitute the foundation of numerical estimation. I showed for the first time that symbolic-number mapping, a basic numerical ability, contributes to numerical estimation, alongside domain knowledge which has been the primary focus of previous research. Through integrating insights from numerical cognition with research on numerical estimation, I thus broaden theoretical perspectives on the factors underlying numerical real-world estimation. In Manuscripts II and III, I explored how changes in the underlying resource domain knowledge elicited through the provision of information systematically shape numerical estimates. In Manuscript II (Kreis, Notarbartolo, et al., 2025), I examined how informational interventions—seeding, rules and labels—that target an updating of specific aspects of domain knowledge can enhance estimation accuracy and promote sustainable decisions. With this work, I demonstrate the applied value of cognitive theories of estimation for applied contexts and underscore the potential of seeding for fostering informed real-world decisions. In Manuscript III (Kreis, Hermann, et al., 2025), I addressed a further consequence of the provision of information in numerical estimation: hindsight bias (e.g., Hawkins & Hastie, 1990; Tversky & Kahneman, 1974). Approaching hindsight bias through the lens of numerical estimation research, I replicate and extend evidence that hindsight bias emerges as a by-product of the same knowledge-updating that also improves estimation accuracy. These findings underscore the value of bridging theoretical traditions to deepen our understanding of numerical estimation and its associated phenomena.

Taken together, through integrating insights from separate traditions and combining ecological validity with methodological rigor, this work advances a more unified understanding of the factors shaping numerical real-world estimation. In the following sections of the discussion, I will further elaborate on the implications of this integrative perspective and outline promising directions for future research and integration.

4.1 Implications and Outlook

The unifying theme across all three manuscripts of this dissertation is the focus on mental resources and their role in shaping numerical real-world estimation. This focus has the potential to serve as a guiding framework in the pursuit of a more integrated perspective on estimation, offering new insights into how numerical estimation can be both theoretically understood and practically supported. Especially the considerations on the malleability of numerical estimation reflecting changes in the underlying mental resources are particularly informative. On the one hand, it can provide valuable theoretical insights into the processes underlying the formation of numerical estimates and, in doing so, offer approaches for effectively supporting these processes. On the other hand, it can help reveal connections to other phenomena involving altered numerical estimates that might otherwise remain unnoticed.

When it comes to improving numerical estimation, previous research has primarily focused on domain knowledge and strategies to enhance estimation through enhancing this resource. However, our findings in Manuscript I, highlighting symbolic-number mapping as an additional mental resource, suggest a complementary pathway. Although the specific mechanisms through which basic numeric abilities influence estimation remain to be fully understood, targeting symbolic-number mapping seems particularly promising: unlike targeting domain-specific knowledge, enhancing symbolic-number mapping may offer more generalizable benefits across a wide range of contexts and estimation domains. To foster symbolic-number mapping, prior research has identified several promising strategies, such as providing corrective feedback on number-line tasks or using worked examples prompting participants to actively reflect on why a given mapping is accurate or inaccurate (Fitzsimmons et al., 2023; Opfer & Siegler, 2007; Thompson & Opfer, 2016). Such training might enhance estimation accuracy through two complementary ways. It might directly support estimation by helping individuals define more precise metrics and more accurately determine the relative position of the target value. Long term, as better symbolic-number mapping accuracy is also related to better numerical memory (Peters & Bjalkebring, 2015; Thompson & Siegler, 2010), it might further support the development of more accurate and detailed knowledge over time. With stronger symbolic-number mapping skills, individuals may integrate new numerical information more effectively into existing knowledge structures and retain more specific and relevant details—resources that can later be drawn upon in estimation tasks.

Whereas the idea of improving estimation accuracy by targeting symbolic-number mapping remains a hypothesis requiring further investigation, efforts to enhance estimation accuracy through strengthening domain knowledge have already been extensively studied and shown to be highly effective (e.g., Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024; Wohldmann & Healy, 2020). Hereby, our research clearly demonstrates the value of understanding improvements in estimation as a consequence of changes in the underlying knowledge. Especially recognizing the distinct roles of metric and mapping knowledge (Brown & Siegler, 1993) can facilitate the identification of knowledge gaps and enable a tailoring of interventions. Looking ahead, a key question will be whether the improvement in knowledge gained from such interventions will translate into changes in real-world behavior. Our findings in Manuscript II, showing improved performance in two lab selection tasks, offer a first indication of practical relevance. Still, future research will need to continue exploring whether informational interventions that improve estimation through knowledge updating also translate into improved decision-making in everyday contexts.

However, changes in domain knowledge shape numerical estimation not only by improving estimation accuracy but also by eliciting hindsight bias. Our findings in Groß et al. (2023) and Manuscript III reveal that knowledge updating underlies both phenomena. This perspective challenges prior assumptions about the conditions under which the bias is elicited. Whereas earlier research attributed hindsight bias to exposure to the correct answer to a specific previous judgment (e.g., Blank et al., 2007; Christensen-Szalanski & Willham, 1991; Roese & Vohs, 2012), our results suggest that it arises more broadly whenever knowledge structures relevant for the corresponding judgment are revised (Groß et al., 2023; Kreis, Hermann, et al., 2025). Supporting this broader view, von der Beck et al. (2019) found that hindsight bias in event-outcome judgments emerged not only when participants learned the actual outcome, but also when they revised assumptions based on new contextual cues.

Our findings also offer a fresh perspective on the question of how problematic hindsight bias truly is. Traditionally viewed as a cognitive distortion that impairs learning (e.g., Biais & Weber, 2009; Fischhoff, 1975), our results point to a more nuanced picture. We observed clear learning effects (i.e., improved estimation accuracy) even in the presence of hindsight bias, suggesting that in contexts in which learning involves acquiring and applying new knowledge, hindsight bias does not necessarily impede learning (Groß et al., 2023; Kreis, Hermann, et al., 2025).

However, in situations in which learning rather depends on recognizing prior uncertainty or reflecting on discrepancies between initial judgments and actual outcomes, hindsight bias may still be detrimental (Anderson et al., 1993; Biais & Weber, 2009; Danz, 2020; Giroux et al., 2016; Harley, 2007).

Building on our integrative perspective on hindsight bias as a phenomenon of numerical estimation being shaped by changes in its underlying knowledge base, our findings also suggest a potential strategy for its mitigation. Although hindsight bias has proven generally resistant to correction (e.g., Arkes et al., 1988; Davies, 1987; Guilbault et al., 2004; Pohl & Hell, 1996; Roese & Vohs, 2012), one promising approach is to help individuals retain access to their initial beliefs (e.g., Hell et al., 1988; Roese & Vohs, 2012). For instance, prompting people to articulate the reasoning behind their initial judgments has been shown to reduce hindsight bias (Hell et al., 1988). In the context of numerical estimation, it may therefore be useful to encourage individuals to explicitly document their metric and mapping assumptions before receiving corrective information. Such a strategy, though still to be empirically tested, could offer a promising avenue for promoting a more accurate recall of ones naive estimates.

Taken together, the preceding reflections highlight the value of integrating diverse perspectives for advancing our understanding of numerical real-world estimation. Further advancing a unified understanding will require continued efforts to connect diverse strands of research, establish a robust empirical base, and develop more comprehensive theoretical frameworks. The following chapter outlines some key challenges and promising directions for pursuing this integration.

4.2 Future Directions of Integration

Achieving an integrated understanding of numerical estimation will be far from trivial. One fundamental challenge already lies in the differing terminologies used across different subfields. For instance, what is referred to as numerical estimation in cognitive psychology (e.g., Brown & Siegler, 1993; Groß et al., 2024; Kreis, Groß, & Pachur, 2025) is often labeled as quantitative judgment or quantitative estimation in the judgment and decision-making literature (e.g., Izydorczyk & Bröder, 2023; von Helversen & Rieskamp, 2009), and sometimes even as perception of numerical values in applied environmental and sustainability research (e.g., Attari, 2014; Attari et al., 2010; Marghetis et al., 2019).

Another challenge and source of fragmentation lies in the differing quantification traditions. Although estimation accuracy is usually assessed as the deviation between an estimate and the actual value, the specific measures vary widely, from absolute and signed deviations (e.g., Groß et al., 2024; Kreis, Hermann, et al., 2025), relative and standardized indices (e.g., Pohl, 2007; Wohldmann, 2015), to logarithmic transformations (e.g., Bröder et al., 2023; Brown & Siegler, 1996). Integrating these approaches and selecting the most appropriate measure is not a trivial task. As illustrated by the different quantifications used across the manuscripts in this dissertation, it can be challenging to choose a measure that both aligns with the statistical properties of real-world data and allows for comparability across studies. Nonetheless, whereas these choices can affect statistical outcomes and should be carefully considered, they are unlikely to obscure broader result patterns when methodological rigor is maintained.

Advancing an integrated understanding of numerical estimation also requires addressing the theoretical fragmentation across the field, both in terms of overarching frameworks and underlying process assumptions. Theoretical models vary widely in form, ranging from descriptive verbal accounts, such as the metrics and mapping framework (Brown & Siegler, 1993), to formal mathematical models as commonly used in cue-utilization research (e.g., Hoffmann et al., 2019; Izydorczyk & Bröder, 2023; Juslin & Persson, 2002). A key task for future work will be to reconcile these approaches and harness their respective strengths: the ecological validity and conceptual depth of verbal models, and the precision and testability of formal models. At the process level, divergence is equally pronounced. Assumptions span from automatic to deliberate mechanisms, and from shallow adjustments (e.g., Tversky & Kahneman, 1974) to deep knowledge revision (e.g., Groß et al., 2023). Clarifying these assumptions and empirically testing competing claims will be critical for bridging this theoretical divide.

Unfortunately, this fragmentation makes it significantly more difficult to arrive at an integrated perspective, as it can obscure conceptual overlaps, shared assumptions and associated phenomena. It may thus be useful to shift the focus away from differences in terminologies, quantifications and theories and instead consider the core dimensions that structure research on numerical estimation. One productive approach may be to organize the field around two key conceptual questions: (1) What are the mental resources shaping numerical estimates? and (2) Which factors and mechanisms underlie altered numerical estimates?

When it comes to the mental resources underlying numerical estimation, a key challenge lies in developing a more complete understanding of the full spectrum of these resources. Manuscript I examined two such resources, basic numeric abilities and domain knowledge, but many open questions remain. For instance, does the impact of those resources differ depending on context? How do these resources interact? And how can we integrate insights on the role of different aspects of knowledge? Whereas this dissertation focused on metric and mapping knowledge as two key aspects of domain knowledge (Brown & Siegler, 1993), the literature highlights a broader range of knowledge types involved in numerical estimation, such as memory of exemplars or knowledge of abstract cue-criterion relationships (e.g., Griffiths & Tenenbaum, 2006; Hoffmann et al., 2019; Izydorczyk & Bröder, 2021; Juslin et al., 2003). Integrating such cue-based approaches into numerical real-world estimation research, however, poses a challenge. It requires the researcher to have detailed knowledge about the cue structure of a given domain, which is rarely the case in real-world contexts (Izydorczyk & Bröder, 2024). Nonetheless, promising approaches have begun to bridge this gap. Murray and Brown (2009) and Bröder et al. (2023), for example, demonstrated that in domains where cues reliably relate to a criterion (e.g., a car's brand to its price; Murray & Brown, 2009), participants do draw on such cues for estimation and their use of cues can be enhanced through seeding. For less clearly structured domains, promising research by Izydorczyk and Bröder (2024) offers a novel method for deriving hidden cue structures through applying pairwise similarity ratings and so-called multidimensional scaling. Examining how cue knowledge shapes numerical real-world estimation may open new perspectives on the aspects of real-world knowledge individuals draw upon.

In addition to identifying the mental resources shaping estimation, it is crucial to also integrate research on the factors that alter estimates. As shown throughout this dissertation, a particularly powerful influence is informational input, as it alters the knowledge base underlying estimation. Whereas this work focused on the deliberate provision of correct information, real-world information is often incidental, varied in credibility, and context-dependent. And prior research suggests these properties matter: more credible, relevant, and intentional information generally has stronger effects on estimates (e.g., Bonaccio & Dalal, 2006; Chapman & Johnson, 1999; Tversky & Kahneman, 1974; Yaniv, 2004). For example, the advice-taking literature has demonstrated that advice from more credible sources, such as labeled experts, affects estimates more strongly than that of less credible sources

(e.g., Bonaccio & Dalal, 2006; Yaniv, 2004). Similarly, in anchoring research, more relevant or plausible anchors tend to exert greater influence (e.g., Chapman & Johnson, 1999; Furnham & Boo, 2011; Tversky & Kahneman, 1974). Moreover, beyond informational input, other factors influencing estimation should also be considered. Contextual and motivational factors, in particular, are likely to play important roles. For example, judgment and decision-making research has shown that situational conditions, such as time pressure or working memory load, can affect cognitive processing and thereby shape judgments (e.g., Bröder & Schiffer, 2006; Hammond, 2000; Pachur & Hertwig, 2006; Rieskamp & Hoffrage, 2008). Motivational factors may also be critical. Individuals who are more motivated to be accurate, may engage more deeply with information, potentially amplifying or moderating its effects (e.g., Epley & Gilovich, 2016; Kunda, 1990; Molden, 2012; Pelham & Neter, 1995). Future frameworks of numerical real-world estimation should therefore consider how and through which processes informational, contextual, and motivational factors jointly shape numerical estimation.

Finally, in all efforts toward integration, it is crucial to stay grounded in the real-world focus that defines this field of research. As Brunswik (1955) emphasized, research on cognitive abilities must consider the ecological contexts in which the mental processes occur. To achieve this, future research will need to systematically examine numerical estimation across a wide range of real-world contexts. Only through such systematic, ecologically valid work can we build theories that truly capture the complexity of numerical estimation in everyday life.

4.3 Conclusion

With this dissertation, I contribute to a more comprehensive understanding of the factors shaping numerical real-world estimation. In line with the notion that scientific progress builds on previous work (Newton, 1675), I demonstrated how integration across research traditions can advance the field, both conceptually and practically. Looking ahead, I outlined promising avenues for further integration. The hope is that continued integration efforts will eventually pave the way for a comprehensive theory of numerical real-world estimation, one capable of both theoretically explaining behavior and guiding approaches to support better-informed decisions. With this in mind, I am hopeful that my work will be a part of the metaphorical giants whose shoulders others will stand on to see further.

5 Bibliography

Anderson, J. C., Lowe, D. J., & Reckers, P. M. J. (1993). Evaluation of auditor decisions: Hindsight bias effects and the expectation gap. *Journal of Economic Psychology*, *14*(4), 711–737. https://doi.org/10.1016/0167-4870(93)90018-G

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*(2), 305–307. https://doi.org/10.1037/0021-9010.73.2.305

Attari, S. Z. (2014). Perceptions of water use. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(14), 5129–5134. https://doi.org/10.1073/pnas.1316402111

Attari, S. Z., DeKay, M. L., Davidson, C. I., & Bruine de Bruin, W. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16054–16059. https://doi.org/10.1073/pnas.1001509107

Bajželj, B., Richards, K. S., Allwood, J. M., Smith, P., Dennis, J. S., Curmi, E., & Gilligan, C. A. (2014). Importance of food-demand management for climate mitigation. *Nature Climate Change*, *4*(10), 924–929. https://doi.org/10.1038/nclimate2353

Bayen, U. J., Erdfelder, E., Bearden, J. N., & Lozito, J. P. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1003–1018. https://doi.org/10.1037/0278-7393.32.5.1003

Bellana, B., Mansour, R., Ladyka-Wojcik, N., Grady, C., & Moscovitch, M. (2021). The influence of prior knowledge on the formation of detailed and durable memories. *Journal of Memory and Language*, *121*, Article 104264. https://doi.org/10.1016/j.jml.2021.104264

Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 378–391. https://doi.org/10.1037/a0021971

Biais, B., & Weber, M. (2009). Hindsight bias, risk perception, and investment performance. *Management Science*, *55*(6), 1018–1029. https://doi.org/10.1287/mnsc.1090.1000

Blank, H., Musch, J., & Pohl, R. F. (2007). Hindsight bias: On being wise after the event. *Social Cognition*, *25*(1), 1–9. https://doi.org/10.1521/soco.2007.25.1.1

Boels, L., Boels, A., Alberto, R., & Hoogland, K. (2025). Citizens' data-ing with contemporary data in their daily life. *ZDM – Mathematics Education*, *57*(1), 87–101. https://doi.org/10.1007/s11858-025-01665-4

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Advances in psychology* (Vol. 54, pp. 75–114). North-Holland. https://doi.org/10.1016/S0166-4115(08)62171-8

Bröder, A., Dülz, E., Heidecke, D., Wehler, A., & Weimann, F. (2023). Improving carbon footprint estimates of food items with a simple seeding procedure. *Applied Cognitive Psychology*, *37*(3), 651–659. https://doi.org/10.1002/acp.4060

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making*, *19*(4), 361–380. https://doi.org/10.1002/bdm.533

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. *Psychology of Learning and Motivation*, *41*, 321–359. https://doi.org/10.1016/S0079-7421(02)80011-1

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534. https://doi.org/10.1037/0033-295X.100.3.511

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review*, *3*(3), 385–388. https://doi.org/10.3758/BF03210766

Brown, N. R., & Siegler, R. S. (2001). Seeds aren't anchors. *Memory & Cognition*, *29*(3), 405–412. https://doi.org/10.3758/BF03196391

Brunswik, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Camilleri, A. R., Larrick, R. P., Hossain, S., & Patino-Echeverri, D. (2019). Consumers underestimate the emissions associated with food but are aided by labels. *Nature Climate Change*, *9*(1), 53–58. https://doi.org/10.1038/s41558-018-0354-z

Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An individual difference analysis. *Journal of Personality*, *51*(4), 605–620. https://doi.org/10.1111/j.1467-6494.1983.tb00868.x

Castells, M. (1997). An introduction to the information age. *City*, *2*(7), 6–16. https://doi.org/10.1080/13604819708900050

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, Activation, and the Construction of Values. *Organizational Behavior and Human Decision Processes*, *79*(2), 115–153. https://doi.org/10.1006/obhd.1999.2841

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*(1), 93–131. https://doi.org/10.1023/A:1005272027245

Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*(1), 147–168. https://doi.org/10.1016/0749-5978(91)90010-Q

Dallacker, M., Hertwig, R., & Mata, J. (2018). Parents' considerable underestimation of sugar and their child's risk of overweight. *International Journal of Obesity*, *42*(5), 1097–1100. https://doi.org/10.1038/s41366-018-0021-5

Danz, D. (2020). Never underestimate your opponent: Hindsight bias causes overplacement and overentry into competition. *Games and Economic Behavior*, *124*, 588–603. https://doi.org/10.1016/j.geb.2020.10.001

Davies, M. F. (1987). Reduction of hindsight bias by restoration of foresight perspective: Effectiveness of foresight-encoding and hindsight-retrieval strategies. *Organizational Behavior and Human Decision Processes*, *40*(1), 50–68. https://doi.org/10.1016/0749-5978(87)90005-7

Dehn, D. M., & Erdfelder, E. (1998). What kind of bias is hindsight bias? *Psychological Research*, *61*(2), 135–146. https://doi.org/10.1007/s004260050020

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, *2*(2-3), 109–140. https://doi.org/10.1080/135467896394474

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. https://doi.org/10.3389/fpsyg.2015.00621

Egnell, M., Galan, P., Farpour-Lambert, N. J., Talati, Z., Pettigrew, S., Hercberg, S., & Julia, C. (2020). Compared to other front-of-pack nutrition labels, the Nutri-Score emerged as the most efficient to inform Swiss consumers on the nutritional quality of food products. *PLOS ONE*, *15*(2), Article e0228179. https://doi.org/10.1371/journal.pone.0228179

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465–485. https://doi.org/10.1037/0033-295X.86.5.465

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, *30*(3), 133–140. https://doi.org/10.1257/jep.30.3.133

Ercin, A. E., & Hoekstra, A. Y. (2014). Water footprint scenarios for 2050: A global analysis. *Environment International*, *64*, 71–82. https://doi.org/10.1016/j.envint.2013.11.019

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 387–414. https://doi.org/10.1037/0278-7393.24.2.387

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fitzsimmons, C. J., Morehead, K., Thompson, C. A., Buerke, M., & Dunlosky, J. (2023). Can feedback, correct, and incorrect worked examples improve numerical magnitude estimation precision? *The Journal of Experimental Education*, *91*(1), 20–45. https://doi.org/10.1080/00220973.2021.1891009

Food and Agriculture Organization of the United Nations. (2015). *2050: Water supplies to dwindle in parts of the world, threatening food security and livelihoods*. Retrieved January 20, 2025, from https://www.fao.org/newsroom/

detail/2050-Water-supplies-to-dwindle-in-parts-of-the-world-threatening-food-security-and-livelihoods/en

Food and Agriculture Organization of the United Nations. (n.d.). *Water use*. Retrieved January 20, 2025, from https://www.fao.org/aquastat/en/overview/methodology/water-use

Friedman, A., & Brown, N. R. (2000). Reasoning about geography. *Journal of Experimental Psychology: General*, *129*(2), 193–219. https://doi.org/10.1037/0096-3445.129.2.193

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

García-González, Á., Achón, M., Carretero Krug, A., Varela-Moreiras, G., & Alonso-Aperte, E. (2020). Food sustainability knowledge and attitudes in the Spanish adult population: A cross-sectional study. *Nutrients*, *12*(10), Article 3154. https://doi.org/10.3390/nu12103154

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Giroux, M. E., Coburn, P. I., Harley, E. M., Connolly, D. A., & Bernstein, D. M. (2016). Hindsight bias and law. *Zeitschrift für Psychologie*, *224*(3), 190–203. https://doi.org/10.1027/2151-2604/a000253

Goldstein, W. M., & Hogarth, R. M. (1997). Judgment and decision research: Some historical context. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 3–65). Cambridge University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Groß, J., & Bayen, U. J. (2015). Adult age differences in hindsight bias: The role of recall ability. *Psychology and Aging*, *30*(2), 253–258. https://doi.org/10.1037/pag0000017

Groß, J., Kreis, B. K., Blank, H., & Pachur, T. (2023). Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General*, *11*(152), 3167–3188. https://doi.org/10.1037/xge0001452

Groß, J., Loose, A. M., & Kreis, B. K. (2024). A simple intervention can improve estimates of sugar content. *Journal of Applied Research in Memory and Cognition*, *13*(2), 282–291. https://doi.org/10.1037/mac0000122

Groß, J., & Pachur, T. (2019). Age differences in hindsight bias: A meta-analysis. *Psychology and Aging*, *34*(2), 294–310. https://doi.org/10.1037/pag0000329

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology*, *26*(2-3), 103–117. https://doi.org/10.1080/01973533.2004.9646399

Gupta, A., Smithers, L. G., Harford, J., Merlin, T., & Braunack-Mayer, A. (2018). Determinants of knowledge and attitudes about sugar and the association of knowledge and attitudes with sugar intake among adults: A systematic review. *Appetite*, *126*, 185–194. https://doi.org/10.1016/j.appet.2018.03.019

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*(4), 255–262. https://doi.org/10.1037/h0046845

Hammond, K. R. (2000). *Judgments under stress*. Oxford University Press.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press.

Harley, E. M. (2007). Hindsight bias in legal decision making. *Social Cognition*, *25*(1), 48–63. https://doi.org/10.1521/soco.2007.25.1.48

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*(3), 311–327. https://doi.org/10.1037/0033-2909.107.3.311

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition*, *16*(6), 533–538. https://doi.org/10.3758/BF03197054

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory*, *11*(4-5), 357–377. https://doi.org/10.1080/09658210244000595

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hoekstra, A. Y. (2017). Water footprint assessment: Evolvement of a new research field. *Water Resources Management*, *31*(10), 3061–3081. https://doi.org/10.1007/s11269-017-1618-5

Hoekstra, A. Y., Chapagain, A. K., Aldaya, M. M., & Mekonnen, M. M. (2012). *The water footprint assessment manual: Setting the global standard*. Routledge.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. https://doi.org/10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2019). Testing learning mechanisms of rule-based judgment. *Decision*, *6*(4), 305–334. https://doi.org/10.1037/dec0000109

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 566–581. https://doi.org/10.1037/0278-7393.26.3.566

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review*, *28*(5), 1495–1513. https://doi.org/10.3758/s13423-020-01861-1

Izydorczyk, D., & Bröder, A. (2023). Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical Bayesian approach. *Decision*, *10*(4), 347–371. https://doi.org/10.1037/dec0000195

Izydorczyk, D., & Bröder, A. (2024). What is the airspeed velocity of an unladen swallow? Modeling numerical judgments of realistic stimuli. *Psychonomic Bulletin & Review*, *31*(3), 1–15. https://doi.org/10.3758/s13423-023-02331-0

Izydorczyk, D., Kreis, B. K., Kilb, M., & Bröder, A. (2025). *# Knowledge Using social media for improving food-related knowledge: A seeding intervention* [Manuscript submitted for publication], Department of Psychology, University of Mannheim.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. https://doi.org/10.1207/s15516709cog2605_2

Kreis, B. K., Groß, J., & Pachur, T. (2025). Real-world estimation taps into basic numeric abilities. *Psychonomic Bulletin & Review*, *32*(3), 1217–1230. https://doi.org/10.3758/s13423-024-02575-4

Kreis, B. K., Hermann, A., Pachur, T., & Groß, J. (2025). *Hindsight bias through knowledge updating: A conceptual replication of Groß et al. (2023)*. https://osf.io/preprints/psyarxiv/ws9un_v1

Kreis, B. K., Notarbartolo, C., Pachur, T., & Groß, J. (2025). *Improving water-footprint estimates and promoting sustainable food choices: A comparison of three simple interventions*. https://osf.io/preprints/psyarxiv/9z3w5_v1

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

LaVoie, N. N., Bourne, L. E. J., & Healy, A. F. (2002). Memory seeding: Representations underlying quantitative estimations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1137–1153. https://doi.org/10.1037/0278-7393.28.6.1137

Lawson, R., & Bhagat, P. S. (2002). The role of price knowledge in consumer product knowledge structures. *Psychology & Marketing*, *19*(6), 551–568. https://doi.org/10.1002/mar.10024

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969–998. https://doi.org/10.1111/j.1551-6709.2009.01045.x

Lohmann, P. M., Gsottbauer, E., Doherty, A., & Kontoleon, A. (2022). Do carbon footprint labels promote climatarian diets? Evidence from a large-scale field experiment. *Journal of Environmental Economics and Management*, *114*, Article 102693. https://doi.org/10.1016/j.jeem.2022.102693

Marghetis, T., Attari, S. Z., & Landy, D. (2019). Simple interventions can correct misperceptions of home energy use. *Nature Energy*, *4*(10), 874–881. https://doi.org/10.1038/s41560-019-0467-2

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Molden, D. C. (2012). Motivated strategies for judgment: How preferences for particular judgment processes can affect judgment outcomes. *Social and Person-*

ality Psychology Compass, *6*(2), 156–169. https://doi.org/10.1111/j.1751-9004.2011.00424.x

Murray, K. B., & Brown, N. R. (2009). A feature-based inference model of numerical estimation: The split-seed effect. *Acta Psychologica*, *131*(3), 221–234. https://doi.org/10.1016/j.actpsy.2009.05.007

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology*, *103*(4), 689–717. https://doi.org/10.1037/a0029461

Newton, I. (1675). *Letter to Robert Hooke*.

Nieder, A. (2019). *A brain for numbers: The biology of the number instinct*. MIT Press.

Nieder, A. (2020). The adaptive value of numerical competence. *Trends in Ecology & Evolution*, *35*(7), 605–617. https://doi.org/10.1016/j.tree.2020.02.009

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169–195. https://doi.org/10.1016/j.cogpsych.2006.09.002

Pachur, T., & Bröder, A. (2013). Judgment: A cognitive processing perspective. *WIREs Cognitive Science*, *4*(6), 665–681. https://doi.org/10.1002/wcs.1259

Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 983–1002. https://doi.org/10.1037/0278-7393.32.5.983

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Patalano, A. L., Zax, A., Williams, K., Mathias, L., Cordes, S., & Barth, H. (2020). Intuitive symbolic magnitude judgments and decision making under risk in adults. *Cognitive Psychology*, *118*, Article 101273. https://doi.org/10.1016/j.cogpsych.2020.101273

Pelham, B. W., & Neter, E. (1995). The effect of motivation of judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology*, *68*(4), 581–594. https://doi.org/10.1037/0022-3514.68.4.581

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology*, *108*(5), 802–822. https://doi.org/10.1037/pspp0000019

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition*, *25*(1), 14–31. https://doi.org/10.1521/soco.2007.25.1.14

Pohl, R. F. (1998). The effects of feedback source and plausibility of hindsight bias. *European Journal of Cognitive Psychology*, *10*(2), 191–212. https://doi.org/10.1080/713752272

Pohl, R. F., Eisenhauer, M., & Hardt, O. (2003). SARA: A cognitive process model to simulate the anchoring effect and hindsight bias. *Memory*, *11*(4-5), 337–356. https://doi.org/10.1080/09658210244000487

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*(1), 49–58. https://doi.org/10.1006/obhd.1996.0064

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258–276. https://doi.org/10.1016/j.actpsy.2007.05.004

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Schley, D. R., & Peters, E. (2014). Assessing "economic value": Symbolic-number mappings predict risky and riskless valuations. *Psychological Science*, *25*(3), 753–761. https://doi.org/10.1177/0956797613515485

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, *89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Schubbe, D., Scalia, P., Yen, R. W., Saunders, C. H., Cohen, S., Elwyn, G., van den Muijsenbergh, M., & Durand, M.-A. (2020). Using pictures to convey health information: A systematic review and meta-analysis of the effects on patient and consumer health behaviors and outcomes. *Patient Education and Counseling*, *103*(10), 1935–1960. https://doi.org/10.1016/j.pec.2020.04.010

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–250. https://doi.org/10.1111/1467-9280.02438

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education*, *3*(3), 143–150. https://doi.org/10.1111/j.1751-228X.2009.01064.x

Sonnenberg, L., Gelsomin, E., Levy, D. E., Riis, J., Barraclough, S., & Thorndike, A. N. (2013). A traffic light food labeling intervention increases consumer awareness of health and healthy choices at the point-of-purchase. *Preventive Medicine*, *57*(4), 253–257. https://doi.org/10.1016/j.ypmed.2013.07.001

Spencer, J. (1961). Estimating averages. *Ergonomics*, *4*(4), 317–328. https://doi.org/10.1080/00140136108930533

Springmann, M., Godfray, H. C. J., Rayner, M., & Scarborough, P. (2016). Analysis and valuation of the health and climate change cobenefits of dietary change. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), 4146–4151. https://doi.org/10.1073/pnas.1523119113

Thompson, C. A., & Opfer, J. E. (2016). Learning linear spatial-numeric associations improves accuracy of memory for numbers. *Frontiers in Psychology*, *7*, Article 24. https://doi.org/https://doi.org/10.3389/fpsyg.2016.00024

Thompson, C. A., & Siegler, R. S. (2010). Linear numerical-magnitude representations aid children's memory for numbers. *Psychological Science*, *21*(9), 1274–1281. https://doi.org/10.1177/0956797610378309

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

United Nations. (2023). *Summary of proceedings by the president of the general assembly*. Retrieved January 10, 2025, from https://sdgs.un.org/sites/default/files/2023-05/FINAL%20EDITED%20-%20PGA77%20Summary%20for%20Water%20Conference%202023.pdf

von der Beck, I., Cress, U., & Oeberst, A. (2019). Is there hindsight bias without real hindsight? Conjectures are sufficient to elicit hindsight bias. *Journal of*

Experimental Psychology: Applied, *25*(1), 88–99. https://doi.org/10.1037/xap0000185

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 867–889. https://doi.org/10.1037/a0015501

Wilson, S. A., Arora, S., Zhang, Q., & Griffiths, T. (2021). A rational account of anchor effects in hindsight bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 1327–1332. https://escholarship.org/uc/item/5g91n4b5

Winzer, E., Wakolbinger, M., Schätzer, M., Blagusz, K., Rieder, A., Lechleitner, M., & Hoppichler, F. (2021). Impact of a nutrition education programme on free sugar intake & nutrition-related knowledge in fifth-grade schoolchildren. *European Journal of Public Health*, *31*(1), 136–142. https://doi.org/10.1093/eurpub/ckaa219

Wohldmann, E. L. (2015). Planting a seed: Applications of cognitive principles for improving food choices. *The American Journal of Psychology*, *128*(2), 209–218. https://doi.org/10.5406/amerjpsyc.128.2.0209

Wohldmann, E. L., & Healy, A. F. (2020). Learning and transfer of calorie information. *Applied Cognitive Psychology*, *34*(6), 1485–1494. https://doi.org/10.1002/acp.3727

World Health Organization. (2015). *Guideline: sugars intake for adults and children*. Retrieved July 24, 2024, from https://www.who.int/publications/i/item/9789241549028

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Zimdahl, M. F., & Undorf, M. (2021). Hindsight bias in metamemory: Outcome knowledge influences the recollection of judgments of learning. *Memory*, *29*(5), 559–572. https://doi.org/10.1080/09658211.2021.1919144

A Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title *On the Factors Shaping Numerical Real-World Estimation* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby conform the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date:

B Copies of Manuscripts

Psychonomic Bulletin & Review (2025) 32:1217–1230
https://doi.org/10.3758/s13423-024-02575-4

BRIEF REPORT

Real-world estimation taps into basic numeric abilities

Barbara K. Kreis[1] · **Julia Groß**[1] · **Thorsten Pachur**[2,3]

Accepted: 12 August 2024 / Published online: 28 October 2024

Abstract

Accurately estimating and assessing real-world quantities (e.g., how long it will take to get to the train station; the calorie content of a meal) is a central skill for adaptive cognition. To date, theoretical and empirical work on the mental resources recruited by real-world estimation has focused primarily on the role of domain knowledge (e.g., knowledge of the metric and distributional properties of objects in a domain). Here we examined the role of basic numeric abilities – specifically, symbolic-number mapping – in real-world estimation. In Experiment 1 ($N = 286$) and Experiment 2 ($N = 592$), participants first completed a country-population estimation task (a task domain commonly used to study real-world estimation) and then completed a number-line task (an approach commonly used to measure symbolic-number mapping). In both experiments, participants with better performance in the number-line task made more accurate estimates in the estimation task. Moreover, Experiment 2 showed that performance in the number-line task predicts estimation accuracy independently of domain knowledge. Further, in Experiment 2 the association between estimation accuracy and symbolic-number mapping did not depend on whether the number-line task involved small numbers (up to 1000) or large numbers that matched the range of the numbers in the estimation task (up to 100,000,000). Our results show for the first time that basic numeric abilities contribute to the estimation of real-world quantities. We discuss implications for theories of real-world estimation and for interventions aiming to improve people's ability to estimate real-world quantities.

Keywords Real-world estimation · Symbolic-number mapping · Domain knowledge · Number-line task

Introduction

People commonly need to estimate unknown quantities in their daily lives – whether gauging how long it will take to get to the train station or assessing the calorie content of a meal. Correctly understanding and estimating such quantities can be highly relevant: It is important not to miss one's train, and to be able to evaluate whether one's calorie intake is in line with one's nutritional goals. Which mental resources are involved in the estimation of real-world quantities?

To date, theoretical and empirical work on real-world estimation – which has been studied across various domains, including country populations, city-to-city distances, longitudes and latitudes, sugar content of food items, frequency of health risks, the number of people participating in different sports, and tuition fees for U.S. universities (Brown & Siegler, 1993, 1996, 2001; Friedman & Brown, 2000; Groß et al., 2024; Lawson & Bhagat, 2002; Pachur, 2024; Pachur et al., 2013) – has focused primarily on the role of domain knowledge. Domain knowledge refers to any knowledge that a person might have of a given domain (e.g., country populations), including knowledge of both qualitative aspects (e.g., geographical features that indicate uninhabitable land such as deserts) and quantitative aspects (e.g., exemplars such as the population of a specific country, or ordinal relationships between countries; Brown, 2002; Brown & Siegler, 1993; Lawson & Bhagat, 2002).

However, is domain knowledge the only mental resource contributing to the accurate estimation of real-world quantities? The role of more basic numeric abilities, such as symbolic-number mapping, has received considerable attention in the context of laboratory tasks involving quantities (e.g., memory for numbers or preference for monetary lotteries) but has not yet been considered in the context of real-world estimation. Symbolic-number mapping (Schneider et al., 2018; Thompson & Siegler, 2010; Peters & Bjalkebring, 2015; Schley & Peters, 2014) refers to a per-

Barbara K. Kreis
barbara.kreis@uni-mannheim.de

[1] Department of Psychology, University of Mannheim, Mannheim, Germany

[2] School of Management, Technical University of Munich, Munich, Germany

[3] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

son's ability to correctly represent magnitudes proportionally to each other on a mental number line. It is often measured with the number-line task (Siegler & Opfer, 2003), where participants are presented with a blank horizontal line marked only with a range (usually from 0 to 100 or 1000) and asked to map a given number (e.g., 42) onto the line (Schneider et al., 2018; Siegler & Opfer, 2003). Performance in the number-line task develops throughout childhood and shifts from a logarithmic to a more linear mapping (Siegler et al., 2009; Opfer & Siegler, 2007). Symbolic-number mapping has been shown to be associated with performance in various complex numeric tasks. For instance, more accurate symbolic-number mapping is related to a better memory for numbers (Thompson & Siegler, 2010; Peters & Bjalkebring, 2015), to choosing the normatively better option more frequently in a risky choice task (Peters & Bjalkebring, 2015; Patalano et al., 2020), and to trading off money and time more proportionally (Schley & Peters, 2014).

Might symbolic-number mapping also be involved in real-world estimation? To date, the two strands of research have existed independently. Studies on the relationship between symbolic-number mapping and judgment and decision-making have, for the most part, relied on decision-making paradigms in which participants are presented with experimentally designed numeric stimuli (e.g., lotteries, numbers of objects; Patalano et al., 2020; Peters & Bjalkebring, 2015; Schley & Peters, 2014). All the information needed to solve the task is provided by the experimenter. By contrast, research on the estimation of real-world quantities relies on real-world stimuli. Here, participants have to retrieve relevant information learned outside the lab from memory. They generate an estimate by integrating various pieces of numeric and non-numeric information from their real-world knowledge. Due to these profound differences in stimuli and task requirements, it seems difficult to determine whether symbolic-number mapping may also play a role in real-world estimation.

However, on a procedural level, there are indications this might be the case. As described by Brown (2002), a real-world estimate often has to be constructed based on a ballpark notion of the general metric of the objects in a domain, and on assessments of the ordinal position of the objects. Specifically, once a general metric or response range for objects in a domain has been set, "estimates are generated by determining the relative or ordinal value of the target item and selecting a numerical value from the appropriate portion of the range" (p. 326). Arguably, mapping the position of objects onto a metric range recruits abilities similar to those required in symbolic-number mapping tasks (even though symbolic-number mapping, unlike real-world estimation, also requires perceptual skills). However, it has not yet been tested whether the accuracy of real-world estimation is associated with symbolic-number mapping.

If such an association exists, this would be informative for theories on real-world estimation as well as for interventions aiming to improve people's ability to estimate real-world quantities. Specifically, in addition to domain-specific approaches such as conveying quantitative knowledge of a domain (e.g., the calorie content of different food items), interventions could include domain-general approaches that target basic numeric abilities, such as providing feedback on the number-line task (e.g., Fitzsimmons et al., 2023; Opfer & Siegler, 2007; Thompson & Opfer, 2016).

Our goal in this article is to assess whether the accuracy of real-world estimation is associated with symbolic-number mapping. In two experiments, participants estimated the population of various countries (a domain commonly used to investigate real-world estimation; e.g., Brown, 2002; Brown & Siegler, 1993, 1996) and performed a number-line task. Experiment 2 further sought to disentangle the effects of symbolic-number mapping and domain knowledge by examining whether performance in the number-line task predicts estimation accuracy independently of domain knowledge; and tested whether the strength of the association between estimation accuracy and symbolic-number mapping depends on the match between the range of the quantities to be estimated and the range of the numbers to be mapped.

Experiment 1

Method

Experiment 1 was part of a larger study investigating the role of knowledge updating in hindsight bias (Groß et al., 2023, Experiment 2). In a first phase, participants estimated the population of several countries. In a second phase, they received different types of numeric information, depending on the experimental condition to which they were assigned. In a third phase, participants were asked to recall their responses from the initial estimation phase. In a fourth phase, they provided estimates for a new set of countries. In a fifth and final phase, participants completed a number-line task. For the present purposes, only the initial estimation phase and the number-line task are relevant. A more comprehensive description of the full design and results can be found in Groß et al. (2023), Experiment 2. The experiment was not preregistered.

Participants

A sample of 295 participants was recruited via Prolific (www.prolific.co). All participants were native speakers of German, aged 18 to 45 years (M = 27.8 years, 116 women, 178 men, one diverse). The majority of participants were currently working (130) or studying (university or college,

122), 29 participants were looking for work, seven were high-school students, six were apprentices, and one was retired. The median completion time for the entire experiment was 26.4 min and participants were paid a fixed compensation of £5.60.

Material

For the estimation task, we compiled three sets of items consisting of 32 countries each (see Appendix Table 1). Each participant estimated one item set. The sets were assigned to participants such that they were presented with comparable frequency across participants. In the number-line task, participants were presented with 22 numbers ranging from 2 to 938, which they were asked to map onto a 20-cm horizontal line, labeled from 0 on the left to 1000 on the right (see Procedure and design for details).

Procedure and design

The experiment was programmed with lab.js (Henninger et al., 2022) and hosted via Open lab (Shevchenko, 2022). After providing informed consent and demographic information, participants completed an estimation task, in which they estimated the population of 32 countries (see Fig. 1, panel A, for a sample item). Responses smaller than 1000 were not allowed, as they would likely reflect either a misunderstanding of the task or a lack of attention. The country names were presented sequentially and in random order for each participant. Participants took a median time of 8.1 seconds for each item. Finally, participants completed the number-line task, in which they indicated the position of 22 numbers on a horizontal number line with a mouse click (see Fig. 1, panel B for a sample item). The numbers were presented sequentially and in fixed order (486, 122, 34, 163, 56, 754, 725, 366, 147, 2, 938, 606, 78, 818, 246, 722, 18, 738, 150, 5, 179, 100; see Opfer & Siegler, 2007). Responses could be corrected if necessary. Participants took a median time of 5.4 s for each number. At the end of the experiment, participants could provide comments of any kind in an open response field.

Data diagnostics

Several exclusion criteria were applied to ensure good data quality. First, participants were automatically excluded by Prolific if they did not finish the study within a specified time limit (106 min for a study with an estimated finishing time of 40 min). Second, we excluded participants who reported technical problems (two participants), having been considerably distracted during participation (one participant), or having looked up actual population figures during the estimation task (one participant). Third, we excluded all responses that were equal to or larger than the current world population, 8 billion, and thus unrealistic (8 responses, 0.09% of all responses in the estimation task). Fourth, we checked for extremely fast responses (i.e., below 1 s), but there were no such responses for either task. Fifth, we excluded participants whose median order of magnitude error (our measure of estimation accuracy; see Eq. 1 below) in the estimation task exceeded the threefold interquartile range (five participants). Sixth, we excluded responses in the number-line task that exceeded the threefold interquartile range for a given number (42 responses, 0.65% of all responses in the number-line task). Finally, we checked whether more than 20% of number-line task responses were excluded for any given participant; this was not the case. The exclusions resulted in a final sample of 286 participants. The data and the analysis code are available at https://osf.io/34dvr/.

Analytic approach

To quantify accuracy in the estimation task, we calculated the deviation of the estimated country population from the actual country population in terms of the order of magnitude

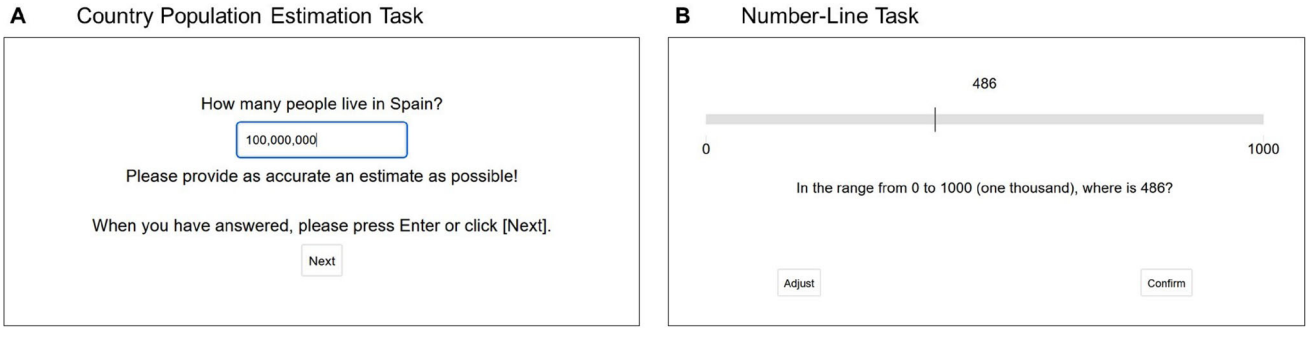


Fig. 1 Sample item from (**A**) the estimation task and (**B**) the number-line task. Translated from German

error (OME) for each item i and for each participant j (Brown & Siegler, 1996):

$$OME_{ij} = |log_{10}\left(\frac{estimate_{ij}}{actual_i}\right)|. \quad (1)$$

The OME is a function of the difference between the estimated and the actual value of an item and converts the difference into an order of magnitude. A larger OME indicates a larger error, that is, lower estimation accuracy. The OME has been frequently used in the context of country-population estimation.[1]

To quantify accuracy in the number-line task, we calculated the relative deviation of the number indicated on the line from the target number for each item k and each participant j:

$$\Delta_{kj} = |\left(\frac{estimate_{kj} - actual_k}{actual_k}\right)|. \quad (2)$$

This measure is commonly used to quantify accuracy in the number-line task, with larger values indicating lower accuracy (Schneider et al., 2018). Each participant's accuracy in the number-line task is calculated as the median Δ across items. A larger Δ indicates lower symbolic-number mapping accuracy.

We used Bayesian linear mixed-effects regression modeling to test whether estimation accuracy (operationalized as the OME) was associated with symbolic-number mapping (operationalized as median Δ). Specifically, the model predicted the OME of participant j for each item i of the estimation task, using as fixed effect the participant's median Δ. The model further included random intercepts for participants and items to take by-person and by-item variability in estimation accuracy into account. The analyses were conducted with the `brms` package (Bürkner, 2017, 2018), which calls `STAN` for MCMC sampling (Stan Development Team, 2019). Prior specification and sensitivity analyses are provided in Appendix B. The general conclusions were robust across different prior specifications.

To statistically evaluate the effects, we compared the full model including a given fixed effect (i.e., symbolic-number mapping), M_1, to the baseline model without that effect, M_0. The baseline model included all random effects that were specified in the full model. We compared the models using the `bayes_factor` function in `brms`, which computes Bayes factors (BF) based on bridge sampling (e.g., Gronau et al. 2017). The BF_{10} quantifies the evidence for the alternative hypothesis relative to the null hypothesis by comparing the full model M_1 to the baseline model M_0.[2]

Results

Participants' average estimation accuracy, measured in terms of median OME across items, was $M = 0.36$ ($SD = 0.19$; range 0.02–1.12). Participants' average symbolic-number mapping accuracy, measured in terms of median Δ, was $M = 0.14$ ($SD = 0.1$; range 0.04–0.68). As Fig. 2 shows, symbolic-number mapping accuracy was associated with accuracy in the estimation task. The results of the mixed-effects model indicated that participants who were more accurate in symbolic-number mapping also provided more accurate estimates for the country populations ($b = 0.38$, $CI_{95\%} = [0.16, 0.59]$).[3]

The standardized regression weight was $\beta = 0.09$, implying that an increase of one standard deviation in symbolic-number mapping accuracy was associated with an increase of 0.09 standard deviations in estimation accuracy. Although the size of the effect may appear modest, evidence for it was strong ($BF_{10} = 83$).[4]

[1] Many real-world domains have highly skewed distributions spanning several orders of magnitude; for such distributions the OME is a better fit than more conventional indices such as the Pearson correlation coefficient or the mean deviation (Brown, 2002; Brown & Siegler, 1996; Groß et al., 2023). In such distributions, a single estimate can differ from others by several orders of magnitude, which would lead to an inflated impact of such outliers if more conventional indices were used. Using a log-based measure such as the OME minimizes the distorting effects of outliers, as under- and overestimation of the same order of magnitude is weighted equally (Brown, 2002). An OME of 1 indicates that the estimation is off by one order of magnitude (e.g., an estimated value of 10 million or 100,000 for an actual value of 1 million); an OME of 2 indicates that the estimation is off by two orders of magnitude (e.g., an estimated value of 100 million or 10,000 for an actual value of 1 million).

[2] The BF is commonly interpreted as follows. A BF_{10} below 1/10 indicates strong evidence, between 1/10 and 1/3 moderate evidence, and between 1/3 and 1 weak evidence for M_0. A BF_{10} larger than 10 indicates strong evidence, between 3 and 10 moderate evidence, and between 1 and 3 weak evidence for M_1 (e.g., Jeffreys, 1998; Lee & Wagenmakers, 2014; van Doorn et al., 2023).

[3] One might object that the association between the performances in the number-line task and the estimation task could be due to individual differences in how diligently participants engaged in solving the tasks. To address this possibility, we conducted additional analyses in which we controlled for how much time participants spent on each of the respective tasks. These analyses showed that the association between symbolic-number mapping and estimation accuracy was independent of the time spent on the tasks. The same held for the results of Experiment 2.

[4] In Appendix C, we report the results for two alternative quantifications of symbolic-number mapping accuracy, namely, the median absolute Δ and the median mapping OME. The general pattern of results was the same for both quantifications of symbolic-number mapping.

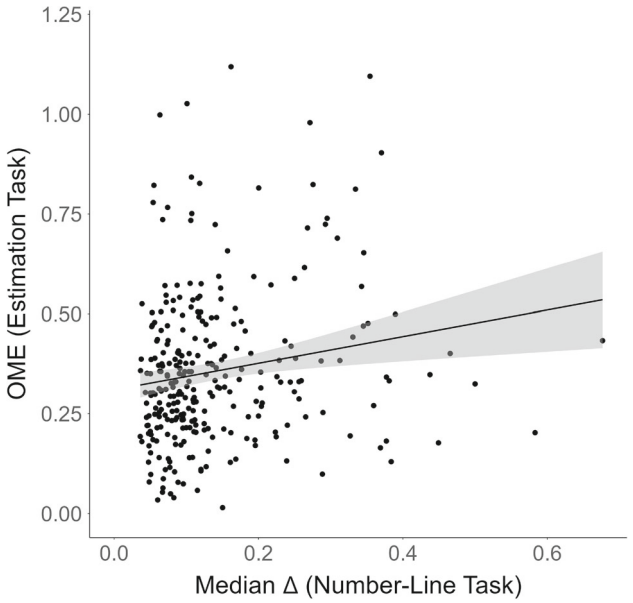


Fig. 2 Association between performance in the estimation task and performance in the number-line task in Experiment 1. OME = order of magnitude error. Each point represents the performance of a participant. For the OME, the median (across items) for each participant is shown. For both OME and median Δ, larger values indicate worse performance. The uncertainty band around the regression line represents the standard error

Discussion

Experiment 1 suggests that basic numeric abilities contribute to real-world estimation. Given the novelty of this finding, replicating it seems desirable. Furthermore, although the statistical evidence for the association was strong, the size of the effect was rather modest. One possible reason is that the number range of the number-line task (up to 1000) did not match that of the estimation task (where the country populations ranged in the millions). The effect size may be larger if the two ranges match. To address this possibility, in Experiment 2 we included two number-line task conditions, one with small (up to 1000) and one with large (up to 100,000,000) quantities. Finally, individuals with a higher symbolic number-mapping ability might also have more knowledge about the domain (e.g., due to a common link with general intelligence). The association observed between symbolic-number mapping and estimation accuracy could thus be due to a confound. We addressed this possibility in Experiment 2 by including a measure of domain knowledge, allowing us to statistically control for domain knowledge.[5]

[5] In the preregistration, this research question was formulated as exploratory. Furthermore, our original formulation of the research question asked whether domain knowledge adds explanatory value to the prediction of estimation accuracy after accounting for symbolic-number mapping. When preparing the analyses, we realized that the other direction is theoretically more reasonable, as prior research has already shown that domain knowledge subserves real-world estimation.

Experiment 2

Method

The experiment was preregistered (see https://aspredicted.org/25K_F7H).

Participants

To determine the target sample size, we conducted a simulation-based a priori power analysis based on the data from Experiment 1 with the `mixedpower` package in R (Kumle et al., 2021). We simulated power for the full model that included symbolic-number mapping (operationalized as median Δ) as a fixed effect. We ran 500 simulations, and defined a critical t value of 2, corresponding to an α level of 5%. The simulation showed that a power of 95% would be achieved with 290 participants. To ensure that our sample size would be sufficient for both individual and joint analyses of the two number-line conditions, we doubled this figure to 580. To account for potential exclusions, we recruited an additional 60 participants, resulting in an overall sample of $N = 640$.

Participants were only allowed to take part in the experiment if they had not taken part in Experiment 1. Of the $N = 640$ participants who submitted their data on Prolific (www.prolific.co), $N = 636$ completed the experiment. The median completion time was 16.7 min and participants received a fixed compensation of £3.70. All participants were native speakers of German, aged 18 to 45 years ($M = 27.9$ years, 277 women, 346 men, 10 diverse, 3 not specified). The majority of participants were working (318) or studying (university or college, 239), 46 were unemployed (33 looking for work, 13 not looking for work), 11 were high-school students, and 22 were apprentices.

Material and design

The estimation task included a fixed set of 46 countries (see Appendix Table 2). We increased the number of items relative to Experiment 1 (which included 32 countries) to increase the power and reliability of the measure. Participants were randomly assigned to one of two conditions of the number-line task. In the *Thousand* condition the number line ranged from 0 to 1000. The task was identical to the number-line task in Experiment 1, with the exception that 18 further numbers were added to achieve a more even distribution across the number range; participants thus mapped a total of 40 numbers (see Procedure for details). In the *Millions* condition the number line ranged from 0 to 100,000,000 and involved the same 40 numbers as in the *Thousand* condition but multiplied by 100,000.

We measured domain knowledge of country populations with a *Domain Engagement Question*. Participants were asked to indicate how often they had engaged with the topic of country populations prior to this study on a seven-point scale. Response options ranged from "very rarely" to "very frequently" (see Appendix D for details). We opted for this approach to measure domain knowledge for several reasons. Asking participants to rate their knowledge directly (Ainley et al., 2002) might tap into self-assessment accuracy and past success at estimating country populations, the latter likely being influenced by a mixture of basic numeric abilities and prior knowledge. Our more indirect approach circumvents this potential problem, focusing on the frequency of engagement with the topic as a purer indicator of the amount of knowledge acquired (e.g., through education and the media).

Procedure

The experiment was programmed with lab.js (Henninger et al. 2022) and hosted via Jatos (Lange et al., 2015). After providing informed consent, participants were asked to estimate the population of 46 countries, one at a time. As in Experiment 1, responses smaller than 1000 were not allowed. Participants took a median time of 7.2 s for each item. Unlike in Experiment 1, the country names were presented in fixed order for all participants. Subsequently, participants performed the number-line task. The procedure and the instructions were the same as in Experiment 1. All 40 numbers were presented sequentially in a fixed order (for the *Thousand* condition: 486, 319, 651, 547, 5, 100, 214, 18, 573, 827, 385, 905, 163, 179, 147, 302, 56, 863, 122, 2, 534, 439, 722, 983, 366, 738, 597, 725, 685, 246, 150, 78, 291, 818, 754, 872, 34, 938, 413, 606). The median response times were 4.3 s and 4.4 s in the *Thousand* and *Millions* conditions, respectively. After completing the two tasks, participants answered the Domain Engagement Question, provided demographic information, and indicated whether they had participated seriously and whether they had cheated (i.e., looked up answers, and/or asked others for help). Finally, participants could provide comments of any kind in an open response field.

Data diagnostics

We preregistered several exclusion criteria. First, participants were automatically excluded by Prolific if they did not finish the study within a specified time limit (71 min for a study with an estimated finishing time of 22 min). For one participant, this exclusion did not work due to technical reasons. The participant had an overall completion time of over 4 h, and was manually excluded by us. Second, we excluded participants who reported problems with the experiment (one participant), having cheated (14 participants), or having just clicked through (three participants). Third, we excluded all responses in the estimation task that were equal to or larger than 8 billion (43 trials, 0.15% of responses in the estimation task). Fourth, we excluded participants whose responses took less than 1 s for more than 10% of items (in both tasks); no such cases occurred. Fifth, we excluded participants whose median OME in the estimation task exceeded the threefold interquartile range (19 participants). Sixth, we excluded responses in the number-line task that exceeded the threefold interquartile range for a given number (separately for the *Thousand* and *Millions* condition); overall, 315 responses were excluded (1.24%). Finally, we excluded participants for whom more than 20% of the responses in the number-line task had to be excluded; this was the case for six participants. The final sample consisted of $N = 592$ participants, with $n = 303$ participants in the *Thousand* condition and $n = 289$ participants in the *Millions* condition.

Analytic approach

As in Experiment 1, we quantified accuracy in the estimation task as the OME (Eq. 1) and accuracy in the number-line task as the median Δ (Eq. 2). To quantify domain knowledge, we converted the responses on the Domain Engagement Question to numbers ranging from 1 (for *very rarely*) to 7 (for *very frequently*).

We used a Bayesian linear mixed-effects regression approach to test for associations of estimation accuracy with symbolic-number mapping and domain knowledge. Specifically, the models predicted the OME of participant j for each item i of the estimation task. As fixed effects, we used the median Δ of participant j, the number-line task condition they were assigned to (*Thousand* vs. *Millions*), as well as their response to the Domain Engagement Question. In addition to the fixed effects, all models included random intercepts for participants and items to take by-person and by-item variability in estimation accuracy into account. Prior specification and sensitivity analyses are described in Appendix B. Again, the general conclusions were robust across different prior specifications. The data and the analysis code are available at https://osf.io/34dvr/.

Results

Is estimation accuracy associated with symbolic-number mapping?

Participants' accuracy in the estimation task was comparable to that of Experiment 1, with a median (across items)

OME of $M = 0.35$ ($SD = 0.19$; range 0.02–1.07), on average. Accuracy in the number-line task measured in terms of median Δ, was $M = 0.07$ ($SD = 0.03$; range 0.02–0.17), on average, and thus better than in Experiment 1 ($M = 0.14$, $SD = 0.1$, range 0.04–0.68). Nevertheless, as Fig. 3 shows, estimation accuracy and accuracy in symbolic-number mapping were again positively associated, as in Experiment 1. Across both number-line conditions, participants with more accurate symbolic-number mapping also did better in the estimation task. This association was corroborated by a mixed-effects model predicting estimation accuracy from symbolic-number mapping ($b = 0.70$, $CI_{95\%}$ = [0.18, 1.21], $BF_{10} = 34$). The standardized regression weight was $\beta = 0.04$.

Does the association depend on the number range used to measure symbolic-number mapping?

Accuracy in the estimation task was similar across the two number-line conditions, with a median (across items) OME of $M = 0.36$ ($SD = 0.19$; range 0.05–1.02), on average, in the *Thousand* condition and $M = 0.35$ ($SD = 0.18$; range 0.02–1.07), on average, in the *Millions* condition. The same held for accuracy in the number-line task, with $M = 0.06$ ($SD = 0.03$; range 0.02–0.17) in the *Thousand* condition and $M = 0.07$ ($SD = 0.03$; range 0.02–0.16) in the *Millions* condition. To test whether the size of the association was larger in the *Millions* condition than in the *Thousand* condition, we compared a model including the interaction of symbolic-number mapping and number-line condition with a baseline model containing only the two fixed effects. Results showed that the association did not differ between the two conditions ($b = -0.03$, $CI_{95\%}$ = [−0.75, 0.69]). While the credible interval of the regression coefficient of the interaction term included zero, the Bayes factor of $BF_{10} = 0.62$ indicated only weak evidence for the absence of an interaction.

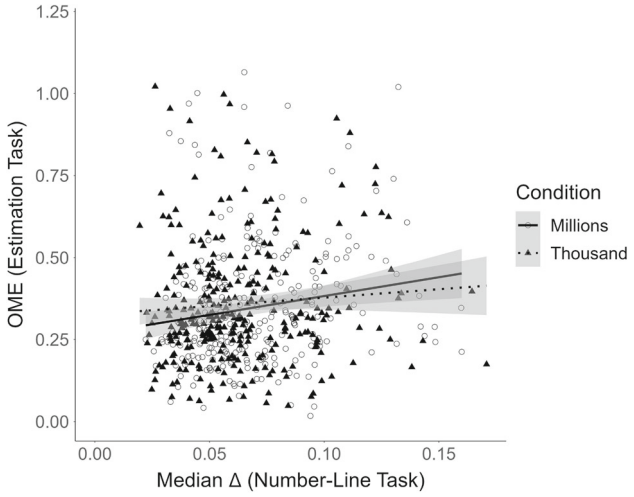


Fig. 3 Association between performance in the estimation task and performance in the number-line task in Experiment 2 by number-line condition. OME = order of magnitude error. Each *point* represents the performance of a participant. For the OME, the median (across items) for each participant is shown. For both OME and median Δ, larger values indicate worse performance. The uncertainty band around the regression line represents the standard error

Does symbolic-number mapping predict estimation accuracy beyond domain knowledge?

Participants' response on the Domain Engagement Question (ranging from 1 = very rarely to 7 = very frequently), was $M = 2.58$ ($SD = 1.37$), on average. Domain knowledge was uncorrelated with symbolic-number mapping ($r = .00$, $BF_{10} = 0.10$), but it was associated with estimation accuracy ($b = -0.05$, $CI_{95\%}$ = [−0.07, −0.04], $BF_{10} > 100,000$). The standardized regression coefficient was $\beta = -0.17$, indicating that an increase of one standard deviation in domain knowledge was associated with an increase of 0.17 standard deviations in estimation accuracy (i.e., a decrease in OME). To test whether estimation accuracy was associated with symbolic-number mapping when statistically controlling for domain knowledge, we compared a model including both domain knowledge and symbolic-number mapping as fixed effects to a baseline model that included only domain knowledge as a fixed effect. Symbolic-number mapping predicted estimation accuracy even when domain knowledge was included as a covariate ($b = 0.71$, $CI_{95\%}$ = [0.24, 1.20], $BF_{10} = 30$). The standardized regression coefficient of symbolic-number mapping was $\beta = 0.04$.[6]

Discussion

We replicated the key finding of Experiment 1, namely, that symbolic-number mapping is associated with accuracy in real-world estimation. The predictive strength of symbolic-number mapping was smaller in Experiment 2 ($\beta = 0.04$) than in Experiment 1 ($\beta = 0.09$). This could be due to the

[6] As in Experiment 1, we also report the results for two alternative quantifications of symbolic-number mapping accuracy in Appendix C, namely the median absolute Δ and the median mapping OME. The general pattern of results was the same for both quantifications of symbolic-number mapping.

overall higher accuracy (i.e., lower median Δ) and smaller variability in symbolic-number mapping in Experiment 2 ($M = 0.07$, $SD = 0.03$, range 0.02–0.17; Experiment 1: $M = 0.14$, $SD = 0.1$, range 0.04–0.68). As Experiment 2 was shorter than Experiment 1, the higher accuracy may be due to higher participant motivation or concentration. Importantly, however, symbolic-number mapping emerged as a robust predictor of real-world estimation accuracy in both experiments, irrespective of differences in accuracy and variability.

Furthermore, the strength of the association between symbolic-number mapping and real-world estimation was not dependent on whether symbolic-number mapping was measured using numbers that matched the range of quantities in the estimation task (country populations in the millions) or not. This indicates that the facet of symbolic-number mapping that is linked to real-world estimation is independent of the range of the numbers used to measure symbolic-number mapping.

Finally, our findings underscore that domain knowledge plays an important role for accurate real-world estimation (Brown & Siegler, 1993; Brown, 2002): Self-reported domain knowledge was a stronger predictor of estimation accuracy ($\beta = -0.17$) than was symbolic-number mapping ($\beta = 0.04$). Crucially, however, symbolic-number mapping predicted estimation accuracy beyond domain knowledge, suggesting a unique role of symbolic-number mapping.

General discussion

Accurately estimating real-world quantities is a relevant skill in daily life. However, empirical and theoretical work on the mental resources that contribute to real-world estimation is scarce. While domain knowledge has been shown to play a role (Brown & Siegler, 1993; Brown, 2002), the potential contribution of basic numeric abilities has not yet been considered. The two experiments reported in this article demonstrated that real-world estimation is reliably associated with symbolic-number mapping. Importantly, we showed that this association is not merely an epiphenomenon arising from a confound between symbolic-number mapping and domain knowledge (e.g., via intelligence): Domain knowledge and basic numeric abilities seem to contribute independently to real-world estimation.

Theoretical implications

Our results can inform and further specify theoretical ideas on real-world estimation. Brown (2002) proposed a conceptual framework outlining the processes that people may engage in when estimating real-world quantities. According to this proposal, people first come up with a general idea of the metric magnitude of the domain in question, such as the range or distribution of objects in that domain. They then locate the relative position of the object to be estimated in this metric range. Where in this process could basic numeric abilities come into play? One possibility is that they influence the estimation process itself: More accurate symbolic-number mapping might allow people to define a more appropriate metric range and to pinpoint the relative position of the object more accurately. Symbolic-number mapping could also facilitate the development of a relevant knowledge base that is recruited for real-world estimation. When presented with numeric information about a domain, people with more accurate symbolic-number mapping might be better able to integrate this information into existing knowledge. Consistent with this idea, people with better symbolic-number mapping have been shown to perform better in memory tasks involving numbers (Thompson & Siegler, 2010; Peters & Bjalkebring, 2015).

Practical implications

Our insights into the mental resources contributing to real-world estimation can inform the development of interventions to boost citizens' ability to estimate real-world quantities. A common approach here is to improve people's domain knowledge by presenting them with a representative selection of items from the corresponding domain (Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024; Marghetis et al., 2019). Although this kind of intervention is simple and effective, the expected improvements are necessarily limited to the specific domain.

Our findings suggest that interventions to boost real-world estimation could also benefit from a more domain-general approach that aims to improve symbolic-number mapping. This could be achieved, for instance, by providing people with corrective feedback on the placement of numbers in a number-line task, or by presenting them with worked examples and asking them to explain why they are correct or incorrect. Prior research suggests that such interventions can improve accuracy not only for symbolic-number mapping itself, but also for other numeric tasks such as memory for numbers (Opfer & Siegler, 2007; Thompson & Opfer, 2016; Fitzsimmons et al., 2023). Improving symbolic-number mapping could improve estimation accuracy by helping people to reorganize their existing knowledge base and correct inaccurate relational ordering or mapping of objects in a domain. In addition, such interventions might help people to integrate numerical information more accurately into the knowledge

base, making subsequent retrieval and use of that information more effective.

Limitations and outlook

In Experiment 2, participants' performance in the number-line task was not affected by whether it involved small or large numbers. This result may seem surprising, given that prior studies reported poorer number-line performance for large numbers (e.g., Landy et al., 2017, 2013). A possible explanation for the similar accuracy in the *Thousand* and *Millions* conditions is that the participants simplified the *Millions* task by mentally crossing out the zeros at the end of the numbers. Future studies could employ numbers without zeros at the end to see whether this affects the results. Note, however, that Landy et al. (2013) and Landy et al. (2017) also used large numbers with zeros at the end and still observed lower mapping accuracy for larger than for smaller numbers. Another possible reason that Landy and colleagues (Landy et al., 2013, 2017) found lower number-line performance for larger than for smaller numbers, whereas we did not, could be that our number-line task with larger numbers only ranged up to 100 million (to match the country populations), whereas it ranged up to a billion in Landy et al. (2013) and Landy et al. (2017).

A potential limitation of our approach to measuring domain knowledge in Experiment 2 is that we used a single self-report item: Participants indicated how often they had engaged with the topic of country populations prior to our study. Arguably, this is a rather indirect measure of domain knowledge. Future studies might consider using genuine knowledge questions about the domain, as in Light et al. (2022). Additionally, future work could include a general measure of cognitive ability, such as IQ, to further clarify the relationship of estimation accuracy with symbolic-number mapping and domain knowledge.

Further, given that our estimation task focused on a specific (though commonly investigated) real-world domain – country populations – it is an open question to what extent our conclusions generalize to other knowledge domains. In general, the relative contribution of symbolic-number mapping and domain knowledge is likely to vary depending on the amount of knowledge available in a domain. For instance, basic numeric abilities might play a smaller role for estimating quantities in familiar domains, such as the sugar content of food items, than in less familiar domains, such as the carbon footprint of food items (Bröder et al., 2023; Groß et al., 2024). Overall, however, we have no basis to assume that the role of basic numeric abilities will disappear completely in familiar domains.

Furthermore, there is a discrepancy between the estimation task and the number-line task regarding the scale format: the estimation task used an open-ended scale, whereas the number-line task used a bounded scale with a clearly defined endpoint. This reflects how these tasks are usually implemented in the literature (for a measurement of symbolic-number mapping with an open-ended scale, see Reinert & Moeller, 2021). Future research could examine to what extent the association between real-world estimation and symbolic-number mapping is affected when both tasks use the same type of response scale (open-ended or bounded).

Conclusion

What role do basic numeric abilities such as symbolic-number mapping play in judgment and decision-making? To date, research has focused primarily on decision-making tasks such as monetary lotteries or trade-offs between money and time, where the numeric stimuli are stated explicitly. Here we found evidence that symbolic-number mapping is also involved in a judgment task where a numeric response has to be constructed by integrating both numeric and non-numeric information from complex real-world knowledge. This highlights the potentially general importance of numeric abilities for judgments requiring numeric responses.

Appendix A: Materials

Materials are shown in Table 1 (Exp. 1) and Table 2 (Exp. 2).

Appendix B: Prior specification and sensitivity analyses

To facilitate specification of the priors, we mean-centered the criterion variable estimation accuracy (operationalized as OME) and the predictors symbolic-number mapping (operationalized as median Δ) and domain knowledge (operationalized as the response to the Domain Engagement Question). We ran prior predictive checks to verify that the priors produced realistic-looking data, as recommended by Schad et al. (2023).

For the intercept parameter (Experiments 1 and 2), we specified a normal distribution `normal(0, 1.5)`. For the standard deviation of the random effects and the residual standard deviation, we specified half-normal distributions, `normal(0, 0.5)` with values > 0.

For the slope parameters, we specified normal distributions with mean zero, implying that they are theory-neutral with regard to potential effects. For the models including

Table 1 Sets of country populations used in Experiment 1

Set 1			Set 2			Set 3		
1	203, 177, 034	Pakistan	**1**	211, 819, 321	Brazil	**1**	268, 501, 680	Indonesia
2	199, 045, 324	Nigeria	**2**	167, 422, 187	Bangladesh	**2**	143, 919, 453	Russia
3	131, 738, 729	Mexico	**3**	107, 505, 862	Philippines	**3**	126, 976, 591	Japan
4	97, 074, 662	Vietnam	**4**	100, 488, 879	Egypt	**4**	109, 159, 044	Ethiopia
5	85, 705, 256	Democratic Republic of the Congo	**5**	82, 592, 416	Turkey	**5**	82, 518, 959	Iran
6	69, 256, 846	Thailand	**6**	59, 246, 609	Italy	**6**	66, 816, 286	United Kingdom
7	60, 229, 204	Tanzania	**7**	57, 812, 482	South Africa	**7**	65, 387, 848	France
8	51, 273, 440	South Korea	**8**	50, 220, 000	Kenya	**8**	54, 158, 522	Myanmar
9	49, 705, 306	Colombia	**9**	46, 439, 538	Spain	**9**	38, 056, 163	Poland
10	45, 169, 147	Uganda	**10**	44, 946, 136	Argentina	**10**	37, 156, 729	Canada
11	43, 877, 093	Ukraine	**11**	42, 425, 837	Algeria	**11**	36, 468, 117	Morocco
12	32, 790, 012	Peru	**12**	40, 002, 380	Iraq	**12**	33, 912, 223	Saudi Arabia
13	32, 294, 009	Malaysia	**13**	36, 883, 979	Afghanistan	**13**	32, 642, 668	Uzbekistan
14	25, 683, 863	North Korea	**14**	31, 404, 292	Angola	**14**	32, 630, 416	Venezuela
15	22, 850, 032	Niger	**15**	29, 858, 634	Ghana	**15**	31, 077, 768	Mozambique
16	20, 106, 983	Burkina Faso	**16**	29, 822, 097	Nepal	**16**	25, 295, 354	Ivory Coast
17	19, 519, 762	Romania	**17**	29,329,832	Yemen	**17**	20, 992, 622	Sri Lanka
18	17, 154, 637	Zimbabwe	**18**	26, 704, 247	Madagascar	**18**	18, 284, 455	Chile
19	17, 114, 912	Netherlands	**19**	25, 074, 109	Cameroon	**19**	17, 938, 326	Zambia
20	14, 600, 000	Somalia	**20**	24, 970, 495	Australia	**20**	17, 452, 973	Guatemala
21	13, 270, 289	Guinea	**21**	19, 510, 631	Malawi	**21**	15, 639, 892	Chad
22	11, 683, 042	Benin	**22**	19, 471, 687	Mali	**22**	13, 132, 406	South Sudan
23	11, 193, 953	Haiti	**23**	18, 518, 517	Kazakhstan	**23**	11, 980, 000	Rwanda
24	11, 133, 944	Greece	**24**	17, 011, 566	Ecuador	**24**	11, 489, 711	Cuba
25	10, 629, 078	Czech Republic	**25**	16, 574, 342	Senegal	**25**	11, 443, 124	Burundi
26	9, 980, 369	Azerbaijan	**26**	16, 394, 043	Cambodia	**26**	11, 430, 000	Tunesia
27	9, 222, 905	Tajikistan	**27**	11, 539, 843	Belgium	**27**	10, 269, 227	Portugal
28	8, 582, 983	Switzerland	**28**	11, 318, 180	Bolivia	**28**	10, 026, 898	Sweden
29	7, 027, 153	Laos	**29**	10, 953, 914	Dominican Republic	**29**	10, 000, 697	Jordan
30	6, 432, 904	El Salvador	**30**	9, 667, 861	Hungary	**30**	9, 439, 781	Belarus
31	4, 833, 127	Ireland	**31**	5, 450, 438	Slovakia	**31**	8, 758, 508	Austria
32	3, 908, 462	Georgia	**32**	4, 149, 214	Croatia	**32**	7, 006, 598	Bulgaria

symbolic-number mapping, condition and their interaction as predictors, we ran prior sensitivity analyses, as Bayes factors can vary depending on the prior distribution; see Schad et al. (2023) or Nicenboim et al. (2021) for discussion. Specifically, we varied the standard deviation of the normal distribution, such that our prior assumptions were differently informed with regard to the expected effect sizes. Table 3 shows which priors were defined for each model and the resulting Bayes factor. In the main text, we report the results for Prior 1. As can be seen, for both experiments, the different priors affected the size of the Bayes factors. However, the general conclusions remained the same.

For the model that contained domain knowledge as the only fixed effect (Experiment 2), we assumed a normal distribution `normal(0, 0.3)` as the prior for the slope parameter.

Appendix C: Results with alternative measures of symbolic-number mapping

In addition to the analyses reported in the main text, we reran the analyses to ensure that the results also hold when using two alternative quantifications. The first is a common

Table 2 Set of country populations used in Experiment 2

Set			Set (continued)		
1	11, 589, 623	Belgium	**24**	31, 072, 940	Ghana
2	31, 255, 435	Mozambique	**25**	32, 971, 854	Peru
3	60, 461, 826	Italy	**26**	206, 139, 589	Nigeria
4	20, 903, 273	Burkina Faso	**27**	69, 799, 978	Thailand
5	9, 449, 323	Belarus	**28**	3, 989, 167	Georgia
6	102, 334, 404	Egypt	**29**	45, 195, 774	Argentina
7	27, 691, 018	Madagascar	**30**	9, 660, 351	Hungary
8	29, 825, 964	Yemen	**31**	59, 308, 690	South Africa
9	97, 338, 579	Vietnam	**32**	8, 654, 622	Switzerland
10	16, 743, 927	Senegal	**33**	51, 269, 185	South Korea
11	11, 402, 528	Haiti	**34**	19, 129, 952	Malawi
12	67, 886, 011	United Kingdom	**35**	10, 203, 134	Jordan
13	11, 673, 021	Bolivia	**36**	10, 847, 910	Dominican Republic
14	212, 559, 417	Brazil	**37**	5, 459, 642	Slovakia
15	11, 193, 725	South Sudan	**38**	45, 741, 007	Uganda
16	109, 581, 078	Philippines	**39**	20, 250, 833	Mali
17	40, 222, 493	Iraq	**40**	17, 915, 568	Guatemala
18	32, 866, 272	Angola	**41**	17, 643, 054	Ecuador
19	43, 851, 044	Algeria	**42**	19, 237, 691	Romania
20	4, 105, 267	Croatia	**43**	16, 718, 965	Cambodia
21	273, 523, 615	Indonesia	**44**	29, 136, 808	Nepal
22	18, 776, 707	Kazakhstan	**45**	84, 339, 067	Turkey
23	46, 754, 778	Spain	**46**	38, 928, 346	Afghanistan

quantification in the symbolic-number mapping literature, the median absolute Δ (Eq. 3; Schneider et al., 2018), the second is a logarithmic measure that is more similar to our logarithmic measure of estimation accuracy, the median mapping OME (Eq. 4).[7] Both are computed as deviation measures for the number indicated (i.e., estimate) from the number presented (i.e., actual) for each item k and each participant j:

$$Absolute\Delta_{kj} = |estimate_{kj} - actual_k|. \quad (3)$$

$$MappingOME_{kj} = |log_{10}\left(\frac{estimate_{kj}}{actual_k}\right)|. \quad (4)$$

Median absolute Δ

For the slope parameter of symbolic-number mapping and its interaction with the number-line condition, we defined as prior a normal distribution `normal(0, 0.00125)` for the median absolute Δ. Overall, the results closely mirrored the results presented in the main text. Estimation accuracy was associated with symbolic-number mapping in both Experiment 1 ($b = 0.0017$, $CI_{95\%} = [0.0002, 0.0032]$, $BF_{10} = 7$, standardized regression weight $\beta = 0.05$) and Experiment 2 ($b = 0.0015$, $CI_{95\%} = [0.0001, 0.0028]$, $BF_{10} = 8$, standardized regression weight $\beta = 0.03$). In addition, there was no interaction between symbolic-number mapping and number-line condition ($b = -0.0003$, $CI_{95\%} = [-0.0021, 0.0017]$, $BF_{10} = 0.79$).

Median Mapping OME

As priors, we used the priors defined in Prior 1 in Appendix B. Again, the results closely mirrored the results presented in the main text. Estimation accuracy was associated with symbolic-number mapping in both Experiment 1 ($b = 0.54$, $CI_{95\%} = [0.20, 0.89]$, $BF_{10} = 41$, standardized regression weight $\beta = 0.07$) and Experiment 2 ($b = 0.77$, $CI_{95\%} = [-0.02, 1.57]$, $BF_{10} = 5.00$, standardized regression weight $\beta = 0.02$). In addition, there was no interaction between symbolic-number mapping and number-line condition ($b = 0.16$, $CI_{95\%} = [-0.72, 1.04]$, $BF_{10} = 0.82$).

[7] Please note that the control analyses with median absolute Δ were preregistered, whereas the analyses with the Mapping OME were not.

Table 3 Prior sensitivity analyses for Experiments 1 and 2

Models	Predictors				Bayes factor
	Symbolic-number mapping	Condition	Symbolic-number mapping × condition	Domain knowledge	
Experiment 1					
Model 1					
Prior 1	**normal(0, 0.5)**	—	—	—	83
Prior 2	**normal(0, 1)**	—	—	—	56
Prior 3	**normal(0, 2)**	—	—	—	28
Prior 4	**normal(0, 5)**	—	—	—	12
Experiment 2					
Model 1					
Prior 1	**normal(0, 0.5)**	—	—	—	34
Prior 2	**normal(0, 1)**	—	—	—	36
Prior 3	**normal(0, 2)**	—	—	—	19
Prior 4	**normal(0, 5)**	—	—	—	5
Experiment 2					
Model 2					
Prior 1	normal(0, 0.5)	normal(0, 2)	**normal(0, 0.5)**	—	0.62
Prior 2	normal(0, 0.5)	normal(0, 2)	**normal(0, 1)**	—	0.62
Prior 3	normal(0, 0.5)	normal(0, 2)	**normal(0, 2)**	—	0.37
Prior 4	normal(0, 0.5)	normal(0, 2)	**normal(0, 5)**	—	0.06
Experiment 2					
Model 3					
Prior 1	**normal(0, 0.5)**	—	—	normal(0, 0.3)	30
Prior 2	**normal(0, 1)**	—	—	normal(0, 0.3)	49
Prior 3	**normal(0, 2)**	—	—	normal(0, 0.3)	30
Prior 4	**normal(0, 5)**	—	—	normal(0, 0.3)	15

Note. Shown in bold is the effect of interest for which the sensitivity analyses were run

Appendix D: Domain Engagement Question

The Domain Engagement Question was worded as follows (translated from German): "At the beginning of this study, you answered several questions about the populations of countries. How *frequently*, prior to this study, have you engaged with this topic (e.g., at school, at work, or in your leisure time)?"

The response options were: very rarely, rarely, rather rarely, neither rarely nor frequently, rather frequently, frequently, very frequently.

Acknowledgements The authors thank Michael Ohlinger for technical support and the programming of Experiment 1 and Susannah Goss for editorial assistance. We further thank Robert Siegler, Ulrich Hoffrage, and Tamara Gomilsek for comments on an earlier version of the article.

Author Contributions Barbara K. Kreis served as lead for software, validation, methodology, investigation, project administration, data curation, formal analysis, visualization, and writing-original draft. Julia Groß served in a supporting role for validation, supervision, methodology, investigation, writing-original draft, and project administration. Thorsten Pachur served as lead for supervision and in a supporting role for validation, methodology, investigation, writing-original draft, and project administration. Barbara K. Kreis, Julia Groß, and Thorsten Pachur contributed equally to design, conceptualization, writing-review and editing, and resources. Julia Groß and Thorsten Pachur contributed equally to funding acquisition.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by Grant GR-4649/4-1 (PA 1925/2-1) from the German Research Foundation (DFG) and supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

Availability of Data and Materials and Code The material for the country-population estimation tasks is provided in Appendix A; the data and analysis code are available via the Open Science Framework https://osf.io/34dvr/.

Declarations

Conflict of Interest The authors have no conflicts of interest to disclose.

Ethics Approval The study protocol for Experiment 1 was approved by the Ethics Committee of the University of Mannheim. This approval was necessary due to a deception manipulation in one of the conditions; this manipulation is not relevant for the present article. No approval was required for Experiment 2 as per the rules of the Ethics Committee of the University of Mannheim.

Consent to Participate Prior to the study, all participants provided informed consent to participate.

Consent for Publication Prior to the study, all participants provided informed consent for publication.

References

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*(3), 545–561. https://doi.org/10.1037/0022-0663.94.3.545

Bröder, A., Dülz, E., Heidecke, D., Wehler, A., & Weimann, F. (2023). Improving carbon footprint estimates of food items with a simple seeding procedure. *Applied Cognitive Psychology, 37*(3), 651–659. https://doi.org/10.1002/acp.4060

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. *Psychology of Learning and Motivation, 41*, 321–359. https://doi.org/10.1016/S0079-7421(02)80011-1

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100*(3), 511–534. https://doi.org/10.1037/0033-295X.100.3.51

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review, 3*(3), 385–388. https://doi.org/10.3758/BF03210766

Brown, N. R., & Siegler, R. S. (2001). Seeds aren't anchors. *Memory & Cognition, 29*(3), 405–412. https://doi.org/10.3758/BF03196391

Fitzsimmons, C. J., Morehead, K., Thompson, C. A., Buerke, M., & Dunlosky, J. (2023). Can feedback, correct, and incorrect worked examples improve numerical magnitude estimation precision? *The Journal of Experimental Education, 91*(1), 20–45. https://doi.org/10.1080/00220973.2021.1891009

Friedman, A., & Brown, N. R. (2000). Reasoning about geography. *Journal of Experimental Psychology: General, 129*(2), 193–219. https://doi.org/10.1037/0096-3445.129.2.193

Groß, J., Kreis, B. K., Blank, H., & Pachur, T. (2023). Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General, 11*(152), 3167–3188. https://doi.org/10.1037/xge0001452

Groß, J., Loose, A. M., & Kreis, B. K. (2024). A simple intervention can improve estimates of sugar content. *Journal of Applied Research in Memory and Cognition*, *13*(2), 282–291. https://doi.org/10.1037/mac0000122

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab.js: A free, open, online study builder. *Behavior Research Methods, 54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Kumle, L., Võ, M.L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods, 53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Landy, D., Charlesworth, A., & Ottmar, E. (2017). Categories of large numbers in line estimation. *Cognitive Science, 41*(2), 326–353. https://doi.org/10.1111/cogs.12342

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science, 37*(5), 775–799. https://doi.org/10.1111/cogs.12028

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE, 10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lawson, R., & Bhagat, P. S. (2002). The role of price knowledge in consumer product knowledge structures. *Psychology & Marketing, 19*(6), 551–568. https://doi.org/10.1002/mar.10024

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* https://doi.org/10.1017/CBO9781139087759

Light, N., Fernbach, P. M., Rabb, N., Geana, M. V., & Sloman, S. A. (2022). Knowledge overconfidence is associated with anti-consensus views on controversial scientific issues. *Science Advances, 8*(29), eabo0038. https://doi.org/10.1126/sciadv.abo0038

Marghetis, T., Attari, S. Z., & Landy, D. (2019). Simple interventions can correct misperceptions of home energy use. *Nature Energy, 4*(10), 874–881. https://doi.org/10.1038/s41560-019-0467-2

Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to Bayesian data analysis for cognitive science. Retrieved February 5, 2024, from https://vasishth.github.io/bayescogsci/book/

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*(3), 169–195. https://doi.org/10.1016/j.cogpsych.2006.09.002

Pachur, T. (2024). The perception of dramatic risks: Biased media, but unbiased minds. *Cognition*, 246, 105736. https://doi.org/10.1016/j.cognition.2024.105736

Pachur, T., Hertwig, R., & Rieskamp, J. (2013). Intuitive judgments of social statistics: How exhaustive does sampling need to be? *Journal of Experimental Social Psychology*, *49*(6), 1059–1077. https://doi.org/10.1016/j.jesp.2013.07.004

Patalano, A. L., Zax, A., Williams, K., Mathias, L., Cordes, S., & Barth, H. (2020). Intuitive symbolic magnitude judgments and decision making under risk in adults. *Cognitive Psychology, 118*, 101273. https://doi.org/10.1016/j.cogpsych.2020.101273

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802–822. https://doi.org/10.1037/pspp0000019

Reinert, R. M., & Moeller, K. (2021). The new unbounded number line estimation task: A systematic literature review. *Acta Psychologica, 219*, 103366. https://doi.org/10.1016/j.actpsy.2021.103366

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2023). Workflow techniques for the robust use of bayes factors. *Psychological Methods, 28*(6), 1404–1426. https://doi.org/10.1037/met0000472

Schley, D. R., & Peters, E. (2014). Assessing "economic value": Symbolic-number mappings predict risky and riskless valuations.

Psychological Science, 25(3), 753–761. https://doi.org/10.1177/0956797613515485

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development, 89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Shevchenko, Y. (2022). Open lab: A web application for running and sharing online experiments. *Behavior Research Methods, 54*(6), 3118–3125. https://doi.org/10.3758/s13428-021-01776-2

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237–250. https://doi.org/10.1111/1467-9280.02438

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education, 3*(3), 143–150. https://doi.org/10.1111/j.1751-228X.2009.01064.x

Stan Development Team. (2019). Stan modeling language: Users guide and reference manual. https://mc-stan.org.

Thompson, C. A., & Opfer, J. E. (2016). Learning linear spatial-numeric associations improves accuracy of memory for numbers. *Frontiers in Psychology, 7*, 24. https://doi.org/10.3389/fpsyg.2016.00024

Thompson, C. A., & Siegler, R. S. (2010). Linear numerical-magnitude representations aid children's memory for numbers. *Psychological Science, 21*(9), 1274–1281. https://doi.org/10.1177/0956797610378309

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2023). Bayes factors for mixed models. *Computational Brain & Behavior, 6*(1), 1–13. https://doi.org/10.1007/s42113-021-00113-2

Improving Water-Footprint Estimates and Promoting Sustainable Food Choices: A Comparison of Three Simple Interventions

Barbara K. Kreis[1], Coralie Notarbartolo[1,2], Thorsten Pachur[3,4], and Julia Groß[1]

[1]University of Mannheim, Germany

[2]University of Ulm, Germany

[3]Technical University of Munich, Germany

[4]Max Planck Institute for Human Development, Berlin, Germany

Author Note

Correspondence concerning this article should be addressed to Barbara K. Kreis, Experimental Psychology Lab, School of Social Sciences, University of Mannheim, L13, 17, Room 512, D-68161 Mannheim, Germany. E-mail: barbara.kreis@uni-mannheim.de

Funding

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences and by Grant GR-4649/4-1 (PA 1925/2-1) from the German Research Foundation (DFG).

Acknowledgments

The authors thank Anita Todd for editing the manuscript.

Data Availability

Data and analysis code are available via the Open Science Framework https://osf.io/9aqbs/.

Declaration of Interests

We have no conflicts of interest to disclose.

CRediT Authorship Contribution Statement

Barbara K. Kreis served as lead for conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, supervision, validation, visualization, and writing original draft. Coralie Notarbartolo served in a supporting role for formal analysis, methodology, software, validation, and investigation. Thorsten Pachur and Julia Groß served in a supporting role for supervision. Barbara K. Kreis and Coralie Notarbartolo served equally in resources. Barbara K. Kreis, Coralie Notarbartolo, Thorsten Pachur, and Julia Groß served equally in writing–review and editing.

Abstract

Water scarcity is intensifying globally, and a major contributing factor is the high water demand of food production. Changing consumers' dietary habits could reduce water use to sustainable levels, but this would require citizens to be well calibrated to the water footprint of food products. In three online experiments (total $N = 411$), we tested three types of interventions to improve people's estimates: *seeding*—presenting actual water-footprint values for selected food products; *rule*—providing a simple rule that distinguishes high- from low-water-consuming foods; and *label*—displaying a relative water-footprint score for selected products. An assessment of people's initial water-footprint estimates (i.e., before the intervention) for a wide range of food products revealed substantial limitations in calibration, both in terms of metric knowledge (e.g., range and average water footprint, which were heavily underestimated) and mapping knowledge (i.e., relative rankings of products). Seeding led to the most comprehensive improvements in people's estimates, enhancing both metric and mapping knowledge. The rule and label interventions improved only mapping knowledge. Importantly, all three interventions also led to an improved ability to choose the most water-efficient food products from multiple options (assessed in a subsequent task). The improved estimation abilities thus translated into more sustainable decisions. Our findings highlight water-footprint calibration as a critical yet underappreciated factor in promoting sustainable consumption. Further, they demonstrate that even brief interventions—particularly seeding—can effectively improve estimation competences and support environmentally responsible behavior.

Keywords: sustainability, water footprint, food production, misestimation, informational interventions, seeding

Introduction

Imagine empty supermarket shelves, skyrocketing food prices, and being unable to take a bath or to water your plants—all due to water scarcity. While this scenario may currently seem dystopian to many, water scarcity and associated effects are projected to affect two-thirds of the global population by 2050 (Food and Agriculture Organization of the United Nations, 2015). Initiatives such as the United Nations (UN) Water Action Decade 2018–2028 (United Nations, 2018), the UN Water Conference 2023 (United Nations, 2023), and the Sustainable Development Goals (United Nations Department of Economic and Social Affairs, 2024) emphasize the urgent need to safeguard water resources and promote sustainable water use.

A particularly important contributor to global water consumption is agriculture, accounting for nearly 70% of freshwater withdrawals (Food and Agriculture Organization of the United Nations, n.d.). Dietary changes could thus significantly reduce water demands, as food products differ greatly in their *water footprint*—the amount of water used in their production. This includes consumptive use (rainwater, surface water, and groundwater) and degradative use (water needed to assimilate polluted freshwater), encompassing all stages from farming, food processing, and retailing to consumption (Hoekstra, 2017; Hoekstra et al., 2012). For example, 1 kg of chicken meat has a water footprint of 4,000 L, whereas 1 kg of tomatoes has a water footprint of only 110 L (Ahrens, 2024). Simulations suggest that reducing global meat and dairy consumption alone could reduce water use to sustainable levels, even with a growing population (Bajželj et al., 2014; Ercin & Hoekstra, 2014; Springmann et al., 2016).

While citizens seem generally motivated to adopt more sustainable diets (Heard & Bogdan, 2021; Sánchez-Bravo et al., 2021; Sanchez-Sabate & Sabaté, 2019; Whittall et al., 2023), a fundamental requirement for sustainable consumption is being able to accurately gauge the environmental impacts of one's food choices (Abrahamse, 2019; Sánchez-Bravo et al., 2021). But are consumers able to correctly assess the water footprint of the foods they consume?

Research on this question is scarce. One study that mainly focused on household water use (which was generally underestimated) also examined people's estimates of a sample of four food items (sugar, rice, cheese, and coffee): People were unable to differentiate the relative water footprints of these products (Attari, 2014). However, as the study used a small set of items and focused exclusively on relative comparisons, it provides only a narrow and incomplete picture of consumer knowledge. Still, there are other indications that the public's understanding of food-related water use may be rather limited. For instance, sustainability research indicates that people have difficulty estimating food-related greenhouse gas emissions and demonstrate limited comprehension of sustainability concepts (Bröder et al., 2023; Camilleri et al., 2019; García-González et al., 2020; Nydrioti & Grigoropoulou, 2023). Together, these findings highlight a need to better understand the scope of consumers' ability to estimate the water footprint of food products and to identify effective interventions for fostering knowledge and the ability to make sustainable food choices.

We addressed these issues by pursuing three objectives. First, to gain a comprehensive understanding of individuals' knowledge of the water footprint of food, we assessed the accuracy of participants' intuitive estimates for a broad range of food products. Second, we evaluated and compared the effectiveness of three different informational interventions in improving estimation accuracy. Third, we examined the extent to which these interventions also promote more sustainable decisions—specifically, selecting the least water-intensive option among multiple food products—thereby testing whether and how well the improved estimation ability is spontaneously applied in a decision context.

The Foundations of Real-World Estimation

To accurately estimate the water footprint of different food products, individuals need an understanding of both the plausible numerical range of water footprints and the relative ordering of food products. Brown and Siegler (1993) described these two fundamental aspects of real-world estimation ability as metric knowledge and mapping knowledge, respectively. *Metric knowledge* refers to an understanding of a domain's

statistical properties, such as the range, central tendency, or distribution of the quantities in a domain, for instance, knowing that the water footprint of a food product typically ranges between 100 and 10,000 L/kg. Inaccurate metric knowledge can lead to systematic misestimations. For instance, Attari (2014) found that household water use was consistently underestimated. If the water footprint of food products is similarly underestimated, it may cause the public to underestimate the true impact of their dietary choices. *Mapping knowledge*, by contrast, pertains to an understanding of the relative ranking of objects within the domain—for example, that vegetables have a smaller water footprint than meat. It allows individuals to compare the objects' relative position on the quantitative dimension. With poor mapping knowledge, people may have difficulty distinguishing between items or have misconceptions about a domain's relative structure. For example, they may mistakenly believe that fruit has a higher water footprint than meat, or that beef has a lower water footprint than chicken. This could hinder sustainable behavior—particularly in everyday contexts where people rely on quick comparisons.

Crucially, sustainable behavior requires both accurate metric and mapping knowledge. Only when individuals understand the overall scale of water use as well as how specific food products are distributed along that scale can they accurately assess the impact of their dietary choices. For example, while mapping knowledge might indicate that meat generally has a higher water footprint than vegetables, combining this with metric knowledge allows individuals to grasp the magnitude of this difference—recognizing that replacing meat with vegetables results in substantially greater water savings than replacing fruit with vegetables.

How well calibrated are people's metric and mapping knowledge of food products' water footprint? Understanding this is crucial to inform interventions and policymaking. Since no studies have examined this question to date, the first aim of the present research was to conduct a comprehensive evaluation of individuals' estimation accuracy in terms of metric and mapping knowledge concerning the water footprint of food products.

Interventions for Improving Real-World Estimation

To the extent that the accuracy of people's intuitive water-footprint estimates is limited, how could the calibration of the estimates be improved? In the following, we describe three interventions that have previously been used to increase people's calibration in real-world estimation and that we implemented in the context of estimates of the water footprint of food products. The second aim of this study was to evaluate and compare these interventions' effectiveness in improving metric and mapping knowledge, and examine whether the acquired knowledge generalized to a broader range of food products, which is crucial for enhancing the interventions' real-world impact.

The first intervention is theoretically grounded in the previously mentioned metrics and mapping framework. In a *seeding* intervention (Brown & Siegler, 1993), participants are provided with *seed facts*—the actual values for a representative subset of items from a domain (e.g., "The water footprint of 1 kg of tomatoes is 110 L"). Several studies have consistently shown that seeding leads to improvements in both metric and mapping knowledge for subsequent estimates of seeded items (Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024; LaVoie et al., 2002; Wohldmann & Healy, 2020). Importantly, the acquired metric knowledge transfers, also improving the estimates of other items of the domain. The transfer of mapping knowledge, by contrast, seems more context-dependent (Bröder et al., 2023; Brown, 2002; Brown & Siegler, 1993, 1996; Groß et al., 2024; Wohldmann & Healy, 2020). For seeding to lead to transfer of mapping knowledge, it seems necessary that people infer from the seed facts generalizable rules about the domain's relative structure. This is particularly likely when the domain consists of distinct subgroups of objects that consistently differ in rank (e.g., Bröder et al., 2023; Brown & Siegler, 1993; Murray & Brown, 2009). In such cases, participants may be able to abstract rules (e.g., vegetables tend to have lower water footprints than meats) and apply them to new items. Moreover, transfer of mapping knowledge through seeding seems more likely when people do not already possess knowledge of the possibly extractable rules. Given that food groups can clearly be ranked in terms of their water footprint and that people have only limited prior

experience with this ranking from their everyday lives, seeding may lead to a transfer of mapping knowledge in this context.

We compared the seeding intervention to two informational intervention strategies stemming from applied sustainability research. In a *rule* intervention, participants are provided with explicit, summarized information about a domain's relative structure, such as subgroup rankings (Marghetis et al., 2019). In the context of the water footprint of food products, such a rule might take the form, "Fruits and vegetables require rather little water, while animal products, beans, nuts, and seeds tend to consume rather a lot of water." Rule interventions are unlikely to improve metric knowledge because they do not convey numerical information. Rather, they provide generalized mapping information that should support the transfer to a broad range of items (e.g., Brown, 2002; Brown & Siegler, 1993; Marghetis et al., 2019). Accordingly, Marghetis et al. (2019) found that the rule "large appliances that primarily heat or cool use a lot more energy than people think they use" (Marghetis et al., 2019, p. 875) enhanced participants' ability to rank appliances' energy consumption through improving their understanding of the appliances' relative energy use.

Finally, in a *label* intervention, participants are provided with simplified, relative ranking information at the item level. The label intervention is frequently applied in health and sustainability contexts (e.g., Brunner et al., 2018; Camilleri et al., 2019; Egnell et al., 2020; Lohmann et al., 2022; Sonnenberg et al., 2013; Watson et al., 2014). A well-known example in Europe is the Nutri-Score, which ranks the nutritional value of food products on a 5-point scale (A to E) relative to other products from the same category (Chantal et al., 2017; Egnell et al., 2020). A water-footprint label could follow a similar approach, assigning relative scores to individual products (e.g., "A banana has a water-footprint score of 2 out of 5"). Like rules, labels do not convey numerical information and thus are not expected to improve metric knowledge, but they should improve mapping knowledge for labeled items. In addition, labels could lead to a transfer of mapping knowledge (i.e., to unlabeled items), as they might help participants infer broader structural patterns of the domain. For instance, if both a

banana and an apple are categorized as Level 2, and eggs and chicken meat are categorized as Level 4, participants may infer that other fruits might also rank lower than other animal-based products.

Overall, the seeding, rule, and label interventions represent different approaches for improving people's estimates of the water footprint of food products, each with distinct strengths in targeting metric or mapping knowledge. Seeding holds particular promise for fostering a broad and transferable ability to estimate water footprints by potentially providing transferable information for both types of knowledge.

Does Improved Estimation Translate Into More Sustainable Decisions?

The third aim of this study was to assess the downstream impact of improved estimation accuracy on decision making. While enhancing metric and mapping knowledge of foods' water footprint is a crucial step, it is ultimately insufficient unless this knowledge can be translated into more sustainable decisions. While rule- and label-based interventions have been shown to support healthier and more sustainable food choices (e.g., Brunner et al., 2018; Camilleri et al., 2019; Edenbrandt & Smed, 2018; Egnell et al., 2020; Lohmann et al., 2022; Marghetis et al., 2019; Sonnenberg et al., 2013; Watson et al., 2014), their impact on decisions that tap into water-footprint knowledge has not yet been systematically assessed. Moreover, it is unexplored which behavioral outcomes seeding can have. We therefore tested whether the three types of interventions helped participants identify the most water-efficient option from several alternatives—allowing us to explore the practical potential of the interventions for supporting more sustainable consumption.

The Present Research

We conducted three experiments in which we assessed participants' level of metric and mapping knowledge of the water footprint of food products and evaluated and compared the impact of the three interventions on knowledge and sustainable decisions. In all experiments, participants first estimated the water footprint of various food items in a pre-intervention estimation task. This served to assess their intuitive estimates and to identify general patterns of over- or underestimation. Participants

were then assigned to one of the intervention groups or to a control group. In the seeding intervention (Experiments 1–3), they received the actual water footprints for all items they had previously estimated. In the rule intervention (Experiment 1), participants were presented with the rule "When comparing the water footprints in the production of different types of food, one can observe: Fruits and vegetables require rather little water, while animal products and beans, nuts, and seeds tend to consume rather a lot of water," explicitly outlining the relative ordering of those food groups. In the label intervention (Experiments 2 and 3), participants were shown labels indicating the relative water footprint of the previously estimated items, with each item assigned to one of five water-footprint levels, from 1 (low water footprint) to 5 (high water footprint).

In the subsequent post-intervention estimation task, participants estimated the water footprint of all previously estimated items as well as of a set of new items to test possible improvements in metric and mapping knowledge and its transfer beyond the specifically learned (i.e., seeded or labeled) items.[1] Afterward, participants completed a selection task to assess whether the interventions led to more sustainable decisions. Specifically, we tested the degree to which participants selected the least water-intensive items among multiple alternatives. Finally, participants filled out three self-report questionnaires assessing prior contact with the topic of the water footprint of food products, general interest in sustainability, and proenvironmental attitude. These measures allowed us to explore whether previous exposure to the topic and environmental attitude were associated with initial estimation accuracy, improvements in knowledge, and sustainable decisions.

In Experiment 1, we compared the effects of the seeding intervention to those of the rule intervention, on both metric and mapping knowledge, as well as on a shopping-list selection task, where participants had to select from four shopping lists

[1] In the rule intervention, improvements for both previously estimated and new items indicate a transfer of the learned rule, as no item-specific information is given and information generally had to be transferred to individual items by design.

the one with the lowest water footprint. In Experiment 2, we compared the effects of the seeding intervention to those of the label intervention. In addition, we modified the selection task such that participants had to select from a set of 20 items the 10 with the lowest water footprint. We wanted the selection task to provide a more flexible decision-making context, focusing on individual items rather than predefined lists. In both Experiments 1 and 2, we examined participants' knowledge of the water footprint of food products for 1 kg of each food item, the standard quantity used in water-footprint calculations and information (Armstrong, 2021; Mekonnen & Gerbens-Leenes, 2020). In Experiment 3 we aimed to replicate Experiment 2 but adjusted the to-be-estimated quantity to 100 g instead of 1 kg per item to examine whether the smaller quantity affected estimation ability. Since 100 g is a more common portion size, it may be easier to conceptualize and could potentially lead to more accurate estimates.

Hypotheses

We expected that all three interventions would enhance participants' ability to estimate the water footprint of food products as well as their selection performance, although to varying degrees. Specifically, we expected the seeding intervention to lead to the most comprehensive improvements—enhancing estimates in terms of both metric and mapping knowledge for previously estimated and new items—and to improve performance in the selection tasks.[2] We expected the rule intervention to improve estimates in terms of mapping (but not metric) knowledge for both previously estimated and new items, and to improve performance in the selection task, although to a lesser degree than for the seeding intervention, which provides item-specific numerical information and thus might facilitate a more accurate comparison of the item lists. Similarly, we expected the label intervention to improve estimates in terms of mapping

[2] Because of mixed previous findings regarding transfer of mapping knowledge, in Experiment 1 we initially expected improvements in mapping knowledge only for seeded items. This expectation was at odds with our hypothesis that selection performance would improve in the seeding group, as such improvement would require transfer of mapping knowledge—a point we had overlooked.

(but not metric) knowledge for both previously estimated and new items, and to improve performance in the selection task.

Method

In the following, we present the methods for Experiments 1, 2, and 3 together, as they are largely identical. Any specific differences are noted where applicable. All experiments were preregistered (Experiment 1 see https://osf.io/67y24; Experiment 2 see https://osf.io/4hjpn; Experiment 3 see https://osf.io/n3cra). Data and analysis code are available via https://osf.io/9aqbs/.

Participants

We determined the target sample sizes for the experiments using a priori simulation-based power analyses with the `mixedpower` package in R (Kumle et al., 2021). For each power analysis, we ran 2,000 simulations, using a critical t value of 2 (α = 5%) and aiming for a power of 80%. For Experiment 1, we based the simulation of the effect of a seeding intervention on metric knowledge for new items, compared to a control group, in the domain of country populations (Experiment 2 in Groß et al., 2023). To account for uncertainty regarding the expected effect size, we simulated power for the lower boundary (in absolute terms) of the 50% credible interval (CI) of the effect. This resulted in a sample size of 105 participants for Experiment 1. For Experiments 2 and 3, we based our simulation on the results of Experiment 1. Specifically, we simulated power for multiple variables, quantifying improvements in both metric and mapping knowledge for new items for the seeding intervention versus the control group.[3] This resulted in a sample size of 120 participants to ensure sufficient power for all variables. For each experiment, we aimed for an additional 21 participants to account for exclusions, resulting in a target sample size of 126 for Experiment 1 and

[3] The simulations for the main quantification of metric knowledge, the order of magnitude error (OME; see Equation 2), resulted in a target sample size of 21 participants. For an alternative quantification of metric knowledge, the absolute deviation between estimated and actual values, the target sample size was 120 participants. For the rank-order correlation ρ, the quantification of mapping knowledge, a sample size of 47 was determined.

a target sample size of 141 for Experiments 2 and 3.

The experiments were conducted online. Participants were recruited via personal contacts, SurveyCircle (2025), and, in Experiment 2, additionally via Sona Systems (https://www.sona-systems.com/). Participation was restricted to individuals aged 18 years or older (for sample size and demographic information see Table 1) and to the use of laptops, desktop computers, and tablets. The use of smartphones was not allowed to ensure an adequate display of the labels. Participation in Experiments 2 and 3 was further restricted to individuals who had not previously participated in any studies on the water footprint of food products. Median completion time was 14.4, 15.0, and 15.8 min for Experiments 1, 2, and 3, respectively. No monetary compensation was awarded for participation, but participants received points they could use to promote their own experiments (SurveyCircle) or course credit (Sona).

Materials

For the water-footprint estimation task, we compiled a list of 50 food items and their associated water footprints in liters per kilogram for Experiments 1 and 2 ($M = 2,983$ L, $SD = 3,452$, $Min = 110$, $Max = 15,547$). For Experiment 3, we scaled the water footprint down to liters per 100 g by dividing the water-footprint values by 10 and rounding to the nearest whole number. The items originated from five food categories—vegetables; fruits; grain products; animal products; and beans, nuts, and seeds—with 10 items per category. The list of items was divided into two sets, A and B, each containing 25 items with five per category, matched on statistical characteristics of their water footprints (mean, median, and range). A list of all food items and their respective water footprints, food groups, and set assignments is provided in Appendix A1.

For the label intervention, we developed a five-level water-footprint label indicating each item's relative water use (see Figure 1). The design drew on the Nutri-Score label (Chantal et al., 2017; Egnell et al., 2020), which classifies foods into five levels (A to E) on the basis of nutritional value and uses a traffic light color scheme ranging from green (A) to red (E). To support an intuitive understanding that our label

Table 1

Demographic Information of Participants in Experiments 1–3

Demographic variable	Experiment 1 $N = 126$	Experiment 2 $N = 146$	Experiment 3 $N = 139$
Age (years)			
Mean age	28.51	24.7	29.0
Age range	18–66	18–61	20–70
Gender			
Female	80	107	95
Male	43	38	43
Nonbinary	1	1	1
Prefer not to disclose	2	-	-
Highest level of education			
GCSE	1	1	3
Completed vocational training	5	2	10
A-levels	44	85	57
University degree	74	57	64
PhD	1	1	4
Prefer not to disclose	1	-	1
Current occupation			
High school	2	-	1
Apprentice	1	-	2
University or college	84	121	87
Working	35	23	47
Not working, looking for work	-	-	-
Not working, not looking for work	3	1	-
Retired	1	1	2

Note. Reported is demographic information of participants before exclusions. GCSE = General Certificate of Secondary Education.

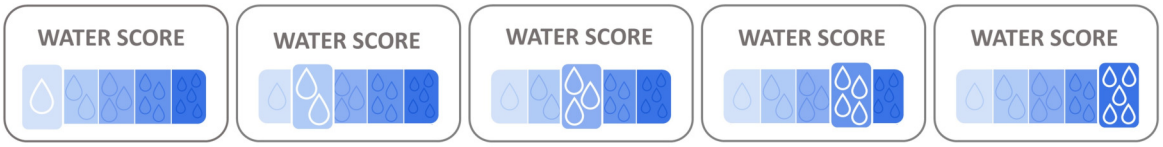
focused on water use, we applied a blue color scheme, with water footprint being represented through a combination of increasing blue saturation and the number of waterdrop icons. For the assignment of food items to label levels, we aimed to create levels of approximately equal size in terms of the number of included items while ensuring that the levels roughly reflected the general ranking of food categories by their average water footprints. For details on the categorization see Appendix A1.

For the shopping-list selection task (Experiment 1), we created four shopping lists of 10 food items each. The lists were composed so as to differ in their average water footprint through varying the proportions of high- and low-footprint items they contained. The lists were assembled from a set of 16 food items, which were compiled in addition to the list of 50 items used in the estimation tasks (details can be found in Appendix A2).

For the food-item selection task (Experiments 2 and 3), we selected a total of 20 food items from the original list of 50 food items, with 10 items from each food set, A and B, and two items per food category per set. The items were selected such that their water footprint was approximately representative of their respective category (see Appendix A1).

Figure 1

Water-Footprint Labels



Note. The box with one drop indicates the lowest water consumption and the box with five drops indicates the highest water consumption.

Additionally, we included three self-report questionnaires to assess participants' prior contact with the topic of the water footprint of food products (in terms of engagement), interest in sustainability, and proenvironmental attitude. To assess prior contact with the water footprint of food products, we asked participants to rate the frequency of their engagement with the topic on a 7-point scale from *very rarely* to *very frequently* ("At the beginning of this study, you answered several questions about the water footprint of food items. How *frequently*, prior to this study, have you engaged with this topic (e.g., at school, at work, or in your leisure time)?"; translated from German). Interest in sustainability was measured with the item "How important is sustainability to you?" (translated from German) rated on a 7-point scale from *very unimportant* to *very important*. Proenvironmental attitude was assessed using the

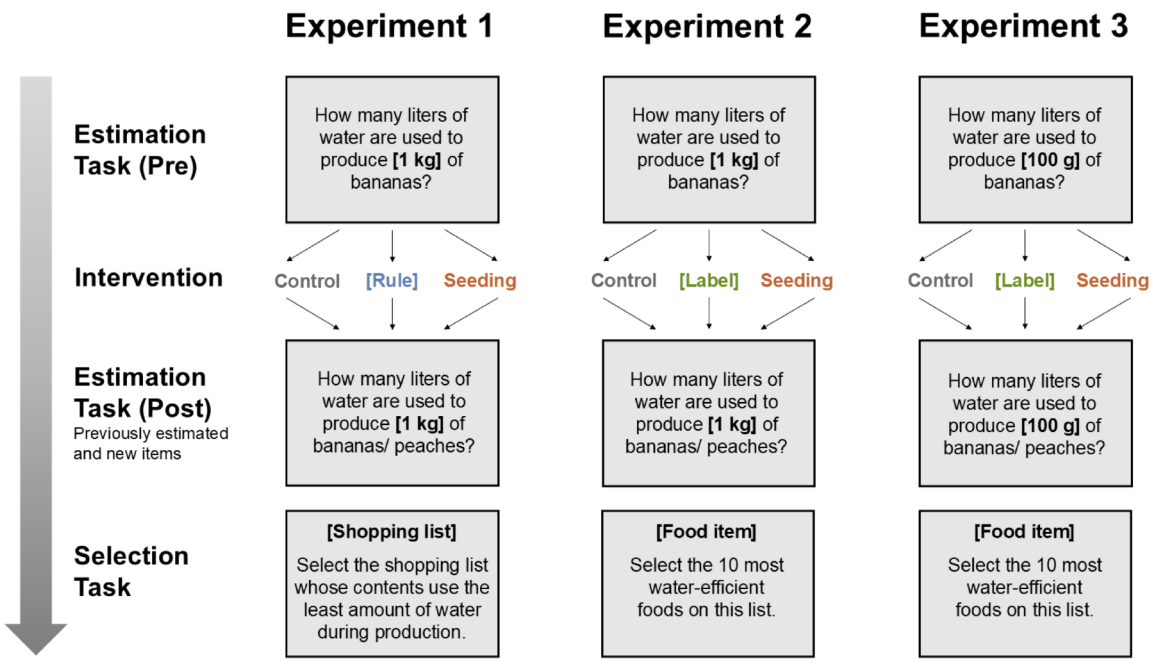
German version of the 15-item New Ecological Paradigm questionnaire, rated on a 5-point scale from *strongly disagree* to *strongly agree* (Marghetis et al., 2019; Schleyer-Lindenmann et al., 2018).

Procedure and Design

The experiments were programmed using lab.js (Henninger et al., 2022) and hosted on JATOS (Lange et al., 2015). The general experimental procedure was the same in each experiment. Participants first provided informed consent and demographic information and then completed the pre-intervention estimation task, followed by the intervention (or a control task), the post-intervention estimation task, and finally the selection task. An overview of the procedure of each experiment is provided in Figure 2.

Figure 2

Procedure and Design of the Experiments



Note. Marked in brackets are aspects that differ between experiments. Estimation Task (Pre) = Pre-intervention estimation task. Intervention = Intervention phase. Estimation Task (Post) = Post-intervention estimation task. Control = Control group. Rule = Rule intervention. Label = Label intervention. Seeding = Seeding intervention. Shopping list = Shopping-list selection task. Food item = Food-item selection task.

In the pre-intervention estimation task, participants were asked to estimate the water footprint of 25 food items (Set A or B), which were presented sequentially and in random order by participant. In the subsequent intervention phase, participants were randomly assigned to one of the intervention groups or the control group. In the *seeding* group, they were shown the actual water footprint values of the 25 previously estimated items (e.g., "To produce 1 kg of bananas, 865 liters of water are used.") in the same random order as in the pre-intervention estimation task, with each item displayed for 5 s, for a total of 125 s. In the *rule* group, participants were presented with the rule (translated from German): "When comparing the water footprint in the production of different types of food, one can observe: Fruits and vegetables require rather little water, while animal products and beans, nuts, and seeds tend to consume rather a lot of water." This rule was displayed for 30 s, followed by an exercise in which participants had to use the rule to classify the four food categories as consuming "rather little" or "rather a lot" of water; they received feedback after each response to help consolidate the rule. The duration of this intervention approximately matched that of the seeding group. In the *label* group, participants were presented with labels showing the relative water footprint of the 25 previously estimated items in the same random order as in the pre-intervention estimation task, each displayed separately for 5 s, for a total of 125 s. In the *control* group, participants read an unrelated text for 125 s.

After the intervention phase, all participants completed the post-intervention estimation task, where they estimated the water footprints of both item sets (25 previously estimated and 25 new items from the other set) in a new random order. Set assignments were counterbalanced across groups.

In the last phase of the experiment, all participants completed the selection task. In Experiment 1, participants were presented with four shopping lists, each containing 10 food items (shopping-list selection task). All four lists were simultaneously displayed in random order for each participant; participants were asked to select the one they believed had the lowest overall water footprint. In Experiments 2 and 3, participants were presented with a list of 20 food items, with the order of food items randomized by

participant (food-item selection task). Participants were instructed to select from these 20 items the 10 they believed had the smallest water footprint.

Participants then completed the three self-report questionnaires (prior contact with the water footprint of food items, interest in sustainability, and proenvironmental attitude). Finally, they indicated whether they had participated seriously and whether they had cheated (e.g., looked up answers or sought outside help) and were given the opportunity to provide additional comments.

This experimental setup resulted for the estimation task in a 2 (Task: pre-intervention vs. post-intervention estimation task) × 2 (Item Type: previously estimated vs. new) × 3 (Group: seeding vs. rule for Experiment 1 vs. label for Experiments 2 and 3 vs. control) mixed design for each experiment. For the selection task, this resulted in a between-subjects design with a three-level factor (Group: seeding vs. rule vs. control for Experiment 1, and seeding vs. label vs. control for Experiments 2 and 3).

Data Diagnostics

Each experiment involved the following preregistered data assessment steps to ensure data quality and exclude noncompliant participants and outliers. First, we excluded participants who exceeded the preregistered 90-min time limit, which was about four times the estimated completion time of 20–25 min ($n = 0$ in Experiment 1, $n = 1$ in Experiment 2, and $n = 0$ in Experiment 3). Second, we excluded participants who reported having cheated ($n = 3$ in Experiment 1, $n = 1$ in Experiment 2, and $n = 2$ in Experiment 3) or just clicked through the experiment ($n = 3$ in Experiment 1, $n = 5$ in Experiment 2, and $n = 7$ in Experiment 3). Third, we excluded participants who reported technical issues ($n = 1$ in Experiment 1, $n = 0$ in Experiment 2, and $n = 0$ in Experiment 3). Fourth, we excluded participants who made only one or two unique estimates across both pre- and post-intervention estimation tasks ($n = 0$ in Experiment 1, $n = 1$ in Experiment 2, and $n = 0$ in Experiment 3). Fifth, we excluded participants who provided more than 10% of their estimates in under 1,000 ms, indicating nonserious task performance ($n = 3$ in Experiment 1, $n = 1$ in Experiment 2, and $n = 5$

in Experiment 3). Sixth, we excluded participants whose median estimation error, the OME (see Equation 2 below), exceeded the threefold interquartile range in the pre- or post-intervention estimation task ($n = 1$ in Experiment 1, $n = 2$ in Experiment 2, and $n = 0$ in Experiment 3). Last, we excluded all remaining individual estimates provided in under 1,000 ms ($n = 6$ in Experiment 1, 0.06% of all estimates; $n = 12$ in Experiment 2, 0.11%; and $n = 16$ in Experiment 3, 0.15%).

The final sample of Experiment 1 thus consisted of 116 participants: 36 in the seeding group, 41 in the rule group, and 39 in the control group. In Experiment 2, the final sample consisted of 136 participants: 45 in the seeding group, 44 in the label group, and 47 in the control group. In Experiment 3, the final sample consisted of 125 participants: 44 in the seeding group, 44 in the label group, and 37 in the control group.

Dependent Variables

As our general measure of metric knowledge as implied by participants' responses in the estimation tasks, we used the OME, which expresses deviations between estimated and actual values in orders of magnitude (Brown & Siegler, 1996).[4] The OME is particularly well-suited to domains with highly skewed distributions where estimated and actual values span several orders of magnitude, such as the water footprint of food items. Compared to conventional indices, such as mean deviation or Pearson correlation, it reduces the distorting impact of outliers.[5]

---

[4] An OME of 1 expresses that the estimated value differs from the actual value by one order of magnitude (e.g., estimating 100 or 10,000 for an actual value of 1,000).

[5] We initially preregistered the absolute deviation between estimated and actual values to quantify metric knowledge in Experiment 1. For Experiments 2 and 3, we preregistered the OME instead, as it better accounts for the highly skewed distribution of food items' water footprints. We had still planned to analyze the absolute deviation as well; however, owing to extreme skewness in the estimates, Bayesian models did not run correctly and bridge sampling did not work for that measure. Thus, we report only the OME, which is more appropriate given the distributions of both actual values and estimates.

To assess initial over- or underestimation, we calculated the signed OME (SOME) for each item i and for each participant j in the pre-intervention estimation task (Brown & Siegler, 1996):

$$SOME_{ij} = log_{10}\left(\frac{estimate_{ij}}{actual_i}\right). \quad (1)$$

Positive values indicate an overestimation of the actual water footprint, and negative values indicate an underestimation of the actual water footprint.

To evaluate improvements in metric knowledge, we calculated the absolute OME in the pre- and post-intervention estimation tasks (Brown & Siegler, 1996):

$$OME_{ij} = |log_{10}\left(\frac{estimate_{ij}}{actual_i}\right)|. \quad (2)$$

For this measure, a smaller OME indicates a smaller error, that is, better metric knowledge.

For the computation of the (S)OME, we transformed all estimates of value 0 to value 1, as estimates of 0 would result in infinite (S)OME values because of the log-based nature of the measure.[6] To quantify participants' mapping knowledge, we calculated the rank-order correlation ρ (Brown & Siegler, 1993) between the estimated and actual values for each participant j separately for the pre- and post-intervention estimation task and for the two item types in the post-intervention estimation task (previously estimated, new). We also computed Fisher's r-to-z transformed rank-order correlation ρ for statistical analyses. Unless otherwise specified, all descriptive statistics and visualizations are based on the raw correlation ρ, while Bayesian analyses used the Fisher-transformed values.

To assess participants' performance in the shopping-list selection task, we calculated the mean water footprint of the selected list for each participant j. To assess participants' performance in the food-item selection task, we calculated the mean water footprint of the 10 selected food items for each participant j. Note that for the shopping-list selection task, we had preregistered that we would analyze the selected

[6] The number of estimates with the value 0 was 12 in Experiment 1 (0.13% of all estimates), 85 in Experiment 2 (0.78%), and 85 in Experiment 3 (0.82%).

list's rank. However, for better comparability with the food-item selection task, we decided to evaluate it using the mean water footprint. Appendix B provides details on the preregistered analyses and a figure depicting list selection by rank. The two quantifications yielded similar results.

To analyze participants' answers in the self-report questionnaires, we converted their responses to numbers, ranging from 1 (for *very rarely*) to 7 (for *very frequently*) for contact with the topic, and from 1 (for *very unimportant*) to 7 (for *very important*) for their interest in sustainability. For their proenvironmental attitude, we calculated a mean score across all 15 items, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*), with higher values indicating a stronger proenvironmental attitude.

Analytic Approach

For all statistical analyses of metric and mapping knowledge in terms of (S)OME and ρ, we applied Bayesian linear mixed-effects models, described below. We employed the `brms` package for parameter estimation (Bürkner, 2017, 2018), which calls `STAN` for Markov chain Monte Carlo sampling (Stan Development Team, 2019). Details on prior specification and sensitivity analyses are provided in Appendix C and Appendix D, respectively. The general conclusions of all three experiments were robust across two alternative prior specifications. To assess whether participants initially over- or underestimated the water footprint of food products, we estimated an intercept-only model, predicting the criterion variable SOME in the pre-intervention estimation task, and we included random intercepts for participants and items.[7]

To analyze whether the interventions improved metric knowledge, we estimated a model predicting OME using task (pre- vs. post-intervention estimation task) as fixed effect, and we included random intercepts and slopes for participants and items. We further compared the size of the improvement in the intervention groups to that in the control group, estimating a model predicting the OME using task, group (control vs. seeding/rule/label), and their interaction as fixed effects, and we included random

[7] While we preregistered for Experiment 1 that we expected a general misestimation, we specified for Experiments 2 and 3 that we expected an underestimation.

intercepts and slopes for participants and items. To analyze how the interventions compared in the size of their associated improvements, we estimated a model predicting the OME using task, group (seeding vs. rule/label), and their interaction as fixed effects, and we included random intercepts and slopes for participants and items.[8]

To assess whether the interventions improved mapping knowledge, we predicted the Fisher's (r-to-z) transformed rank-order correlation ρ using task as fixed effect, and with random intercepts and slopes for participants. We further compared the size of the improvement in the intervention groups to that in the control group, estimating a model that predicted the (r-to-z) transformed rank-order correlation ρ using task, group (control vs. seeding/rule/label), and their interaction as fixed effects, and with random intercepts and slopes for participants. We further analyzed how the interventions compared in the size of their associated improvements, by predicting the (r-to-z) transformed rank-order correlation ρ using task, group (seeding vs. rule/label), and their interaction as fixed effects, and with random intercepts and slopes for participants.[9] To evaluate group differences in participants' performance in the selection tasks, we applied nonhierarchical Bayesian linear regression models, predicting the mean water footprint of the selected list (Experiment 1) or items (Experiments 2 and 3) with group as fixed effect.[10]

We further explored whether participants' self-reported prior contact with the topic, interest in sustainability, and proenvironmental attitude would be associated with initial metric and mapping knowledge, improvements in these measures following the interventions, and performance in the selection tasks. To this end, we extended the previously reported Bayesian regression models to include the self-report variables, with

---

[8] For completeness, we report all group comparison interactions for metric knowledge, although the comparison between the control and the rule/label groups was not preregistered.

[9] For Experiment 1, we report all group comparisons for mapping knowledge, although these were not preregistered. For Experiments 2 and 3, the comparison between the seeding and the rule/label groups was not preregistered.

[10] In the preregistration of Experiment 2, we had erroneously indicated we would include a random intercept for participants, which is not possible with the nonhierarchical data structure.

each self-report measure being examined in a separate model.

For hypothesis testing, we compared the full model including the fixed-effect predictor of interest, M_1, to a baseline model without that predictor, M_0. Each baseline model was defined such that it included all random effects that were specified in the full model. We conducted the model comparisons using the `bayes_factor` function in `brms`, which calculates Bayes factors (BFs) through bridge sampling (e.g., Gronau et al., 2017). The BF_{10} quantifies the evidence for the alternative hypothesis relative to the null hypothesis when comparing the full model M_1 to the baseline model M_0.[11]

Results

How Well Are Participants' Intuitive Estimates of Food Items' Water Footprints Calibrated?

Participants' initial water-footprint estimates in the pre-intervention estimation task revealed a strong underestimation, evident across multiple aspects. For one, the vast majority of individual estimates were underestimates (see "Percentage of types of estimates" in Table 2 for details on percentages of over- and underestimates for each experiment). Further, the median of participants' estimates across items was far below the actual median (see "Median—Actual" and "Median—Estimate" in Table 2). The underestimation was also present at the item level: In Experiments 1 and 2, the median of the estimates per item underestimated the actual water footprint for all 50 items; in Experiment 3, this was the case for 43 out of 50 items. Underestimation was especially pronounced for items with high water footprints (see Figure 3).

The pronounced underestimation was confirmed by a statistical evaluation of participants' initial metric knowledge, as implied by their estimates. An intercept-only model predicting the SOME revealed an estimated intercept of $b = -1.3$, 95% CI $[-1.51, -1.08]$, for Experiment 1, reflecting an average underestimation of 1.3 orders of

[11] Commonly, the BF is interpreted as follows (e.g., Jeffreys, 1961; Lee & Wagenmakers, 2014; Wagenmakers et al., 2018): A BF_{10} below 1/10 is interpreted as strong, between 1/10 and 1/3 as moderate, and between 1/3 and 1 as weak evidence for M_0. Conversely, a BF_{10} above 10 is interpreted as strong, between 3 and 10 as moderate, and between 1 and 3 as weak evidence for M_1.

Table 2

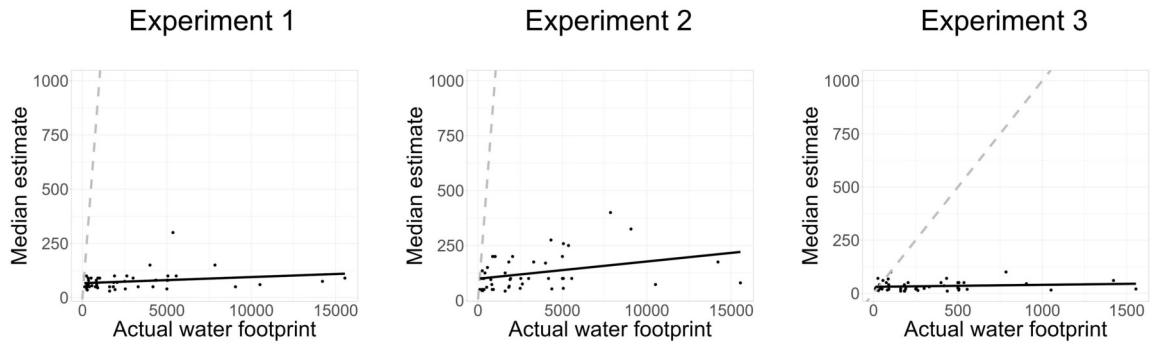
Initial Estimation Accuracy in Experiments 1–3

Statistic	Experiment 1	Experiment 2	Experiment 3
Median			
Actual	1,838	1,838	184
Estimate	70	100	30
Percentage of types of estimates			
Underestimates	89.45%	88.24%	74.92%
Overestimates	10.41%	11.62%	24.79%
Accurate estimates	0.14%	0.15%	0.29%
Rank-order correlation ρ			
Mean	0.20	0.23	0.14
Range	−0.42 to 0.85	−0.53 to 0.74	−0.76 to 0.73

Note. Actual = Actual water footprint in liters. Estimate = estimated water footprint in liters. Accurate estimates = perfect estimate with zero deviation from the actual value.

Figure 3

Initial (i.e., Pre-Intervention) Median Estimates per Item in Experiments 1–3



Note. Each point represents one food item. Shown are, in liters, the actual water footprint per food item on the x axis and the median estimates per food item (across participants) in the pre-intervention estimation task on the y axis. The gray dashed reference line shows where estimates would align with the actual values.

magnitude relative to the actual values. This would correspond to, for example, an estimate of 50.1 L for an actual water footprint of 1,000 L. For Experiment 2, the estimated intercept was $b = -1.15$, 95% CI $[-1.34, -0.96]$, and for Experiment 3, the estimated intercept was $b = -0.67$, 95% CI $[-0.87, -0.45]$.

The accuracy of participants' relative ranking of food items (i.e., mapping knowledge) was also limited. The mean rank-order correlation ρ between the estimated and actual water footprints across participants was 0.20, 0.23, and 0.14 for Experiments 1, 2, and 3, respectively (see "Rank-order correlation ρ" in Table 2 for the range and mean of the rank-order correlation for each experiment).

Participants' lack of knowledge, as evident in their OME and rank-order correlation ρ, was further corroborated by their self-reports, which indicated rather rare prior contact (Experiment 1 = 2.41, Experiment 2 = 2.27, and Experiment 3 = 2.21; on a scale ranging from 1 = *very rare* to 7 = *very frequent*). Hereby, greater self-reported prior contact was associated with better initial metric knowledge, while there was no conclusive evidence regarding an association with mapping knowledge. In contrast, participants' self-reported interest in sustainability and proenvironmental attitude indicated considerable interest (Experiment 1 = 5.53; Experiment 2 = 5.29; Experiment 3 = 5.30, on a scale from 1 = *very unimportant* to 7 = *very important*) and proenvironmental attitude (Experiment 1 = 3.94; Experiment 2 = 3.87; Experiment 3 = 3.85, on a 5-point scale; Schleyer-Lindenmann et al., 2018 ranging from 1 = *do not agree at all* to 5 = *fully agree*); however, neither was associated with metric or mapping knowledge. Furthermore, neither self-report measure was associated with improvements in metric or mapping knowledge. The detailed results for the analyses of the self-reports can be found in Appendix E.

How Well Did the Interventions Improve the Estimates?

In light of participants' limited ability to accurately estimate foods' water footprints, did the interventions improve participants' estimates in terms of metric and mapping knowledge—and how did they fare relative to each other? Results regarding metric knowledge are shown in the left panel of Figure 4, which displays the change in OME from the pre- to the post-intervention estimation task, separately for previously estimated and new items in each group. An improvement in metric knowledge would be indicated by a decrease in OME—that is, an OME difference smaller than 0 (for detailed results including regression coefficients, standardized β coefficients, CIs, and

BFs for all experiments, see Appendix D1). Results in terms of mapping knowledge are shown in the right panel of Figure 4, which displays the change in rank-order correlation ρ from the pre- to the post-intervention estimation task, separately for previously estimated and new items in each group. An improvement in mapping knowledge would be indicated by an increase in rank-order correlation (for detailed results including regression coefficients, standardized β coefficients,CIs, and BFs for all experiments, see Appendix D2).

Effects on Metric Knowledge

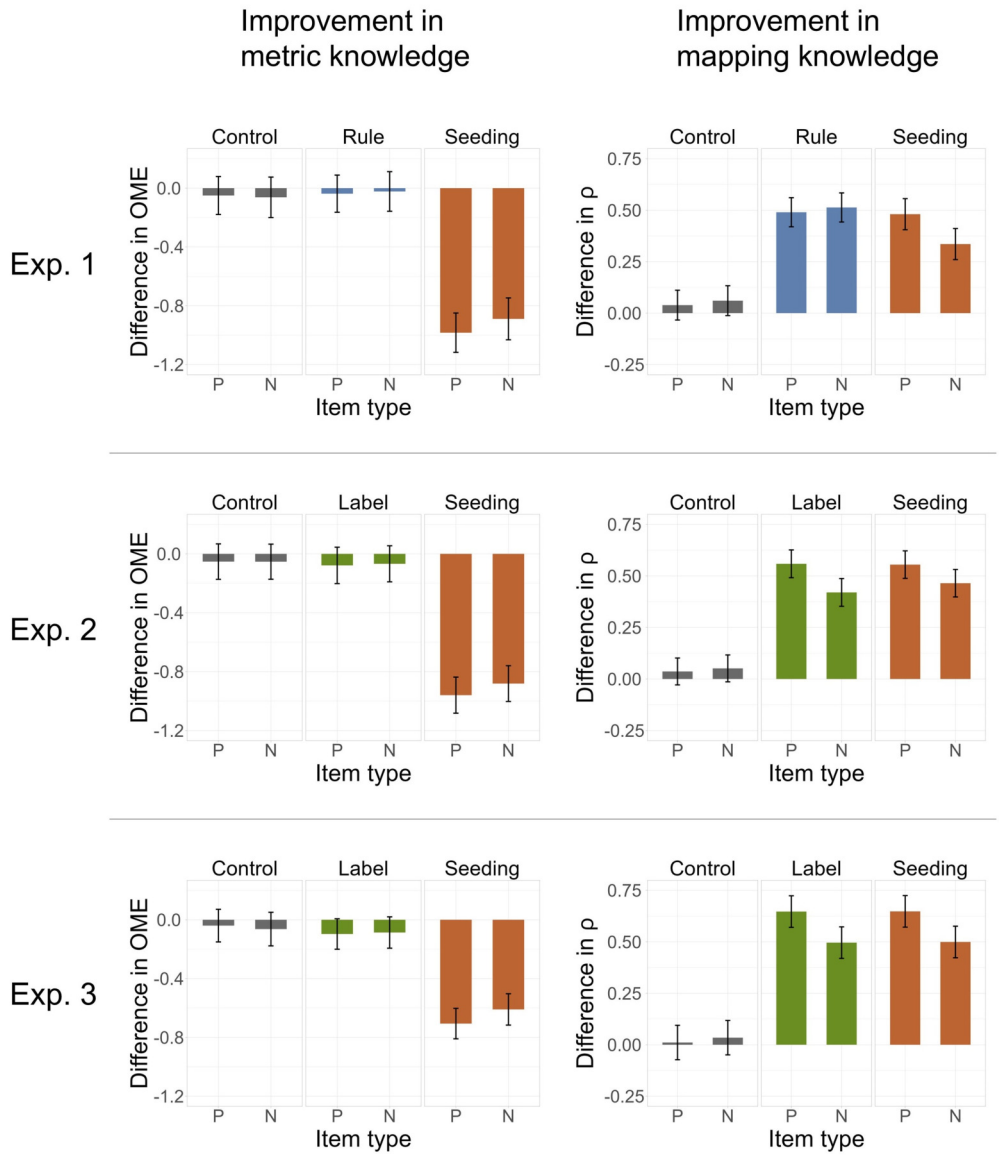
As expected, the seeding intervention yielded strong evidence for an improvement in terms of metric knowledge for both previously estimated items and new items (all $BF_{10} > 10,000$ in all three experiments). Further, these improvements were larger than those in the control group (all $BF_{10} > 10,000$).

In contrast, the rule intervention (Experiment 1) did not lead to an improvement in terms of metric knowledge, neither for previously estimated ($BF_{10} = 0.06$) nor for new items ($BF_{10} = 0.05$). The effects also did not differ from those in the control group (both $BF_{10} \leq 0.07$ for previously estimated and new items). Similarly, for the label intervention in Experiment 2 there was no improvement in terms of metric knowledge, neither for previously estimated ($BF_{10} = 0.17$) nor for new items ($BF_{10} = 0.12$). In Experiment 3, the label intervention again did not improve estimates for new items ($BF_{10} = 0.24$), and the evidence was ambiguous for previously estimated items: While the BF indicated no effect ($BF_{10} = 0.53$), the CI did not include 0, suggesting a small effect ($b = -0.10$, 95% CI $[-0.18, -0.01]$). Importantly, the effects in the label group did not differ from those in the control group (all $BF_{10} \leq 0.10$ for previously estimated and new items, in Experiments 2 and 3).

Interaction analyses further showed that in Experiment 1 improvements in the seeding group were larger than those in the rule group for both previously estimated and new items (all $BF_{10} > 10,000$). Improvements in the seeding group were also larger than those in the label groups across both Experiments 2 and 3, for both previously estimated and new items (all $BF_{10} > 10,000$).

Figure 4

Effects of the Interventions on Metric and Mapping Knowledge as Manifest in the Estimation Tasks



Note. Shown are contrasts between conditional predictions from the corresponding regression model. Specifically, each bar represents the predicted difference in OME/ρ between the pre-intervention estimation task and the post-intervention estimation task. Contrasts are computed separately for previously estimated (P) and new (N) items. Error bars indicate 95% credible intervals. Exp. 1 = Experiment 1, Exp. 2 = Experiment 2, Exp. 3 = Experiment 3. OME = Order of magnitude error. ρ = rank-order correlation. Control = Control group. Rule = Rule intervention. Label = Label intervention. Seeding = Seeding intervention.

Effects on Mapping Knowledge

As expected, there was strong evidence that the seeding intervention improved estimates in terms of mapping knowledge for previously estimated and new items (all $BF_{10} > 10,000$ in all three experiments). Those improvements were larger than those in the control group (all $BF_{10} \geq 2,546$). Similarly, the rule intervention (Experiment 1) yielded strong evidence of improvement for both item types (both $BF_{10} > 10,000$), which also exceeded the effects of the control group (both $BF_{10} > 10,000$). For the label intervention (Experiments 2 and 3), we likewise found strong evidence for improvements in mapping knowledge for previously estimated and new items (all $BF_{10} > 10,000$), which also were larger than the effects found in the control group (all $BF_{10} > 10,000$).

Interaction analyses indicated that the improvements for previously estimated items were similarly large across the three intervention groups (seeding vs. rule: $BF_{10} = 0.28$; seeding vs. label: $BF_{10} \leq 0.20$). For new items, there was weak evidence that the rule intervention led to larger improvements than the seeding intervention ($BF_{10} = 3$). Finally, the seeding and the label interventions led to similar improvements for new items ($BF_{10} \leq 0.21$).

How Well Did the Interventions Improve Sustainable Decisions?

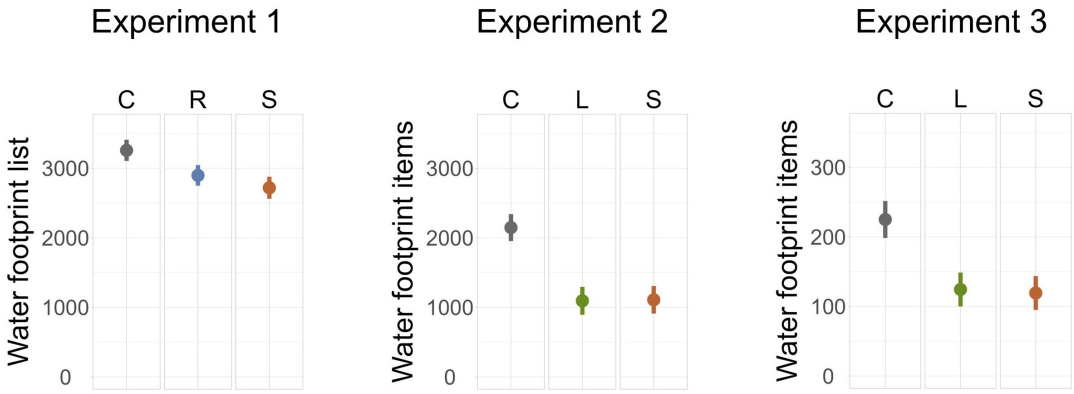
Finally, we examined whether the interventions led to more sustainable decisions, measured as participants' performance in the selection tasks. An improvement would be indicated if the mean water footprint of the selected options in the respective intervention groups is lower than that in the control group. Results are shown in Figure 5.

All three interventions improved performance relative to that in the control group (seeding, Experiment 1: $b = -514.84$, 95% CI $[-719.29, -312.25]$, $\beta = -0.977$[12],

[12] For the selection tasks, the reported partially standardized regression coefficient β reflects the difference in mean water footprint of the selected shopping list (Experiment 1) or food items (Experiments 2 and 3) between the compared groups, expressed in units of the criterion's standard deviation. As the group variable is categorical, the coefficient was standardized with respect to the criterion.

Figure 5

Effects of the Interventions on Sustainable Decisions as Manifest in the Selection Tasks



Note. Shown are the conditional predictions based on the corresponding regression model (estimated means and 95% credible intervals). Water footprint list = Mean water footprint of the selected list in the shopping-list selection task (Experiment 1). Water footprint items = Mean water footprint of the selected food items in the food-item selection task (Experiments 2 and 3). C = Control group. R = Rule intervention. L = Label intervention. S = Seeding intervention.

$\text{BF}_{10} > 10,000$; Experiments 2 and 3: $b = -1027.10$, 95% CI $[-1349.42, -700.32]$, $\beta = -1.11$, $\text{BF}_{10} > 10,000$; $b = -103.97$, 95% CI $[-142.08, -65.69]$, $\beta = -1.03$, $\text{BF}_{10} > 10,000$, respectively; rule, Experiment 1: $b = -342.02$, 95% CI $[-560.58, -124.67]$, $\beta = -0.64$, $\text{BF}_{10} = 22$); label, Experiments 2 and 3: $b = -1039.09$, 95% CI $[-1361.96, -718.18]$, $\beta = -1.10$, $\text{BF}_{10} > 10,000$; $b = -98.52$, 95% CI $[-140.76, -56.16]$, $\beta = -0.91$, $\text{BF}_{10} = 2030$, respectively). Comparing the groups against each other indicated that there was weak evidence that the rule and seeding interventions ($b = -170.35$, 95% CI $[-375.46, 35.86]$, $\beta = -0.36$, $\text{BF}_{10} = 0.80$) and the seeding and label interventions ($b = 14.46$, 95% CI $[-147.75, 173.20]$, $\beta = 0.04$, $\text{BF}_{10} = 0.05$; $b = -5.10$, 95% CI $[-30.04, 19.91]$, $\beta = -0.09$, $\text{BF}_{10} = 0.09$, respectively) led to similar levels of improvement.

Discussion

With water scarcity being alarmingly on the rise, there is a dire need to develop and employ effective strategies to lower global water consumption. Shifting toward less

water-intensive diets is a particularly impactful approach (Bajželj et al., 2014; Ercin & Hoekstra, 2014; Springmann et al., 2016). However, making such dietary changes requires a good ability to estimate the water footprint of food products.

In three experiments, we assessed for the first time participants' intuitive estimates of the water footprint of a broad range of food items. All three experiments indicated considerable inaccuracies in participants' intuitive estimates, reflecting pronounced limitations in people's knowledge in this domain—both in terms of metric knowledge—knowledge of the numerical properties of the water footprint, including its range, central tendency, and distribution—and mapping knowledge—knowledge about the ranks of individual food products relative to each other (Brown & Siegler, 1993). Participants' estimates were characterized by a strong underestimation of food products' actual water footprints. While underestimation was slightly more pronounced when estimates were given in units of liters per kilogram of food (Experiments 1 and 2) than when given in liters per 100 g (Experiment 3), it was present across both quantities, suggesting a general lack of metric calibration rather than a mere difficulty conceptualizing a certain food amount. This aligns with previous research on sustainability-related domains showing that participants underestimate key figures, including household water consumption (Attari, 2014), energy consumption (Camilleri et al., 2019), and greenhouse gas emissions from food production (Bröder et al., 2023; Camilleri et al., 2019). The three experiments further showed that participants' mapping knowledge, in terms of their rank-order correlation, was also alarmingly low across all experiments (Experiment 1: $\rho = .20$, Experiment 2: $\rho = .23$, Experiment 3: $\rho = .14$) and thus notably less accurate than mapping knowledge in other domains, such as country populations ($\rho = .35 - .4$; Brown & Siegler, 1993, 1996), greenhouse gas emissions from food production ($\rho = .58$; Bröder et al., 2023), or sugar content of food products ($\rho = .56$; Groß et al., 2023). These results suggest that participants do not have a well-calibrated intuition of which food products consume more or less water.

Aiming to improve this knowledge, we investigated the impact of three targeted informational interventions: the seeding intervention—based on the metrics and

mapping framework (Brown & Siegler, 1993)—and two interventions originating from applied sustainability research. All interventions—seeding, rule, and label—improved participants' knowledge and supported more sustainable decisions. However, only the seeding intervention led to consistent improvements across all assessed aspects: metric knowledge, mapping knowledge, and sustainable decisions. Specifically, it enhanced both metric and mapping knowledge for both previously estimated (i.e., seeded) and new items, demonstrating a broad transfer beyond the specific facts that were presented. It is worth highlighting that the expected transfer of mapping knowledge did occur—an effect that, while anticipated, had previously been demonstrated only under specific conditions (e.g., Bröder et al., 2023; Murray & Brown, 2009). This may reflect that participants extracted previously unknown, generalized information about the domain's relative structure from the seeded items (as discussed in Bröder et al., 2023; Brown, 2002; Brown & Siegler, 1993). This was likely possible because food categories are fairly predictive of items' relative water footprint, and participants—given their limited prior knowledge—had not derived this rule from earlier exposure. In all three experiments, the seeding intervention further consistently promoted more sustainable decisions in the selection task when compared to the control group and was hereby as effective as the rule and label interventions.

As expected, the rule intervention enhanced only mapping knowledge, but it did so for both previously estimated and new items. This indicates that participants were able to successfully transfer the rule to a wide range of items. Furthermore, they were able to apply this learned mapping information in the decision scenario, selecting less water-intensive lists than the control group. Similar to the rule intervention, the label intervention improved mapping knowledge only—again for both previously estimated (i.e., labeled) and new items. It is likely that participants recognized that items from certain food categories were frequently associated with low or high water-footprint scores, thus enabling a transfer of mapping knowledge beyond the labeled items. Additionally, and as expected, the intervention enhanced sustainable decisions in the selection task, with participants in the label group selecting less water-intensive food

items than the control group.

Practical Implications

Despite increasing research and public discourse on sustainability in recent years (Bettencourt & Kaur, 2011; Science Decade, n.d.; United Nations Department of Economic and Social Affairs, 2024), our findings reveal that knowledge about the water footprint of food products constitutes a critical blind spot that seems to have not yet reached general public awareness. This knowledge gap can present a barrier to sustainable behavior on two levels. First, without knowing which food products are more or less water-intensive, individuals lack the basic prerequisite for making sustainable consumption decisions. Even if people are motivated to adapt their food choices, they cannot engage in more sustainable behavior if they lack knowledge of what constitutes sustainable consumption. This is supported by our findings that across all three experiments, participants' interest in sustainability and proenvironmental attitude were not associated with their initial estimation accuracy. Only participants' self-reported prior contact with the topic—that is, opportunities to acquire knowledge—was associated with the accuracy of their estimates. Second, limited knowledge may reduce motivation to engage with the issue in the first place. Research has shown that accurate knowledge of environmental impacts is associated with greater concern, intention, and engagement in sustainable behaviors (Gugenishvili & Laine-Kronberg, 2025; Ranney & Clark, 2016; Truelove & Parks, 2012). In the context of the water footprint of food products, especially the substantial underestimation may lead individuals to perceive dietary choices as of little environmental relevance, possibly decreasing the likelihood that they will seek further information or be motivated to change their behavior. This highlights the need to make the issue more salient in society and to explicitly target the public's knowledge of the water footprint of food to enable informed and sustainable consumption.

Our study demonstrates that informational interventions—such as seeding, rules, or labels—are an effective means to improve knowledge with minimal cost. Among these, the seeding intervention, grounded in cognitive theory on knowledge acquisition

and estimation accuracy (Brown & Siegler, 1993), emerged as particularly promising. Crucially, the seeding intervention produced combined improvements in both metric and mapping knowledge. This dual enhancement is especially important because, together, the two types of knowledge should enable individuals to better grasp the real-world impact of their dietary choices: Improved metric knowledge helps correct the widespread underestimation of food items' water consumption and conveys an understanding of the scale on which the footprint varies, while mapping knowledge allows consumers to accurately position different food products within this range of water footprints. Moreover, the seeding intervention was as effective as the rule and label interventions in promoting sustainable decisions. These findings illustrate how interventions grounded in cognitive theory, such as seeding, can effectively bridge theoretical insight and practical application, making them a valuable addition to the toolkit for sustainability communication and behavior change.

With its comprehensive impact, the seeding intervention appears particularly well suited to building core knowledge and could be easily applied as an educational approach in classrooms and workshops or through detailed informational materials. However, rule and label interventions also hold important complementary value. For instance, in situations where individuals already possess substantial mapping knowledge from prior experience, seeding alone may contribute little new generalizable ranking information. Conversely, in domains where the structure of item rankings is unclear or complex, rules and labels may more effectively clarify these relationships than a seeding intervention. Moreover, because rules and labels provide information in a simple way, they should represent attractive tools to being used as concise reminders in everyday environments such as billboards, social media, menus, or product packaging, helping support sustainable decision making in daily life. Combining the interventions—using the seeding intervention as a "knowledge starter" and reinforcing it through rules or labels—may thus offer a promising pathway toward comprehensive knowledge gain and supporting sustainable decisions.

Limitations and Outlook

While the seeding intervention proved highly effective and seems particularly well suited to educational settings, its current implementation—presenting individual water footprint values for a rather extensive set of 25 representative food items—may be less practical for everyday contexts. However, several adjustments could enhance its applicability as an everyday intervention. Research suggests that even a smaller, carefully selected set of representative items could improve knowledge (e.g., Groß et al., 2024; LaVoie et al., 2002). Alternatively, presenting summarized numerical information—such as mean and range—could strengthen calibration while remaining brief and accessible (LaVoie et al., 2002). This approach may be even more effective when paired with intuitive reference units (e.g., "5,000 liters of water equals around 33 bathtubs"; see Camilleri et al., 2019, for a similar example). However, it is important to note that such aggregated data can only improve metric but not mapping knowledge.

Another brief but impactful approach could be presenting seed facts incrementally over time. Izydorczyk et al. (2025) tested this for the domain of sugar content and greenhouse gas emissions of food products, delivering two seed facts per day via Instagram. While they observed moderate improvements in metric and mapping knowledge, there was no transfer of mapping knowledge—even for the greenhouse gas emissions of food products, for which a previous study using a conventional seeding intervention did find such an effect (Bröder et al., 2023). This suggests that for mapping transfer to emerge, seed items may need to be presented in ways that support direct comparison and facilitate pattern recognition and rule abstraction. In such contexts, combining seeding with explicit mapping aids—such as rules or labels—may enhance its effectiveness. Camilleri et al. (2019), for example, combined numerical information with explicit mapping information. This could help convey both the scale and structure of the domain more clearly.

Taken together, our findings underscore the potential for further tailoring and developing seeding-based interventions to enhance their usability across diverse settings. By combining different intervention approaches and adapting presentation formats,

these tools can be made more efficient and impactful, expanding their reach beyond educational contexts into broader public communication and behavioral design.

A notable limitation of the present research is that our sampled consisted predominantly of well-educated participants, who also reported a considerable interest in sustainability and a generally proenvironmental attitude. Compared to a more representative sample, their initial knowledge and motivation to engage with the information might be larger, potentially leading to larger improvements in metric and mapping knowledge as well as sustainable decisions. However, our findings suggest that neither factor plays a substantial role in our results. Despite their educational background and eco-conscious mindset, participants demonstrated pronounced inaccuracies in their initial estimates and reported very limited prior contact with the topic of food-related water footprints. Moreover, participants' environmental attitude and interest in sustainability were not reliably associated with either their initial estimation accuracy or their improvements. Nonetheless, future research should further examine the generalizability of these findings to more diverse and less environmentally involved populations.

Finally, it remains unclear whether the more sustainable decisions observed in our experiments would translate into similarly sustainable decisions in everyday life. While prior findings suggest that providing sustainability information can positively influence every-day consumption choices (Camilleri et al., 2019; Gómez-Llanos et al., 2020; Potter et al., 2021; Thorndike et al., 2014; Wardle et al., 2000), the persistent intention–behavior gap remains a challenge, as contextual factors—such as setting, social norms, or time limitations—can impede follow-through despite good intentions (Jekauc et al., 2024; Maher & Dunton, 2020; Papini et al., 2023; Sheeran, 2002). Addressing these barriers remains a key task for sustainability research, exploring how to close the gap between knowing what is sustainable and consistently acting upon it.

Conclusion

Addressing the challenge of sustainable water use requires an integrated approach—one that encompasses not only technological and policy solutions but also insights from the social and behavioral sciences (Pulizzi, 2022; "The Water Action Decade is up for review," 2023). We have highlighted a key yet underexplored psychological factor: public knowledge of the water footprint of food. Our results show that individuals lack accurate knowledge about the water required for food production, but that simple and targeted informational interventions can improve this knowledge effectively and with minimal cost. Enhancing public knowledge can empower individuals to consume more sustainably and complement policy and technological efforts. Our research underscores the essential role of psychological and educational approaches in tackling environmental challenges.

References

Abrahamse, W. (2019). Chapter 3 - Behavior change interventions. In W. Abrahamse (Ed.), *Encouraging Pro-Environmental Behaviour: What Works, What Doesn't, and Why* (pp. 27–45). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-811359-2.00003-2

Ahrens, S. (2024). *Wasserverbrauch bei der Erzeugung von Lebensmitteln 2018*. Retrieved February 10, 2025, from https://de.statista.com/statistik/daten/studie/249969/umfrage/zur-herstellung-von-verschiedenen-nahrungsmitteln-benoetigtes-wasser/

Armstrong, M. (2021). *Which foods need the most water to produce?* Retrieved January 20, 2025, from https://www.weforum.org/stories/2021/06/water-footprint-food-sustainability/

Attari, S. Z. (2014). Perceptions of water use. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(14), 5129–5134. https://doi.org/10.1073/pnas.1316402111

Bajželj, B., Richards, K. S., Allwood, J. M., Smith, P., Dennis, J. S., Curmi, E., & Gilligan, C. A. (2014). Importance of food-demand management for climate mitigation. *Nature Climate Change*, *4*(10), 924–929. https://doi.org/10.1038/nclimate2353

Bettencourt, L. M. A., & Kaur, J. (2011). Evolution and structure of sustainability science. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), 19540–19545. https://doi.org/10.1073/pnas.1102712108

Bröder, A., Dülz, E., Heidecke, D., Wehler, A., & Weimann, F. (2023). Improving carbon footprint estimates of food items with a simple seeding procedure. *Applied Cognitive Psychology*, *37*(3), 651–659. https://doi.org/10.1002/acp.4060

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 41, pp. 321–359). Academic Press.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534. https://doi.org/10.1037/0033-295X.100.3.511

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review*, *3*(3), 385–388. https://doi.org/10.3758/BF03210766

Brunner, F., Kurz, V., Bryngelsson, D., & Hedenus, F. (2018). Carbon label at a university restaurant – Label implementation and evaluation. *Ecological Economics*, *146*, 658–667. https://doi.org/10.1016/j.ecolecon.2017.12.012

Bürkner, P.-C. (2017). Brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Camilleri, A. R., Larrick, R. P., Hossain, S., & Patino-Echeverri, D. (2019). Consumers underestimate the emissions associated with food but are aided by labels. *Nature Climate Change*, *9*(1), 53–58. https://doi.org/10.1038/s41558-018-0354-z

Chantal, J., Hercberg, S., & World Health Organization. Regional Office for Europe. (2017). Development of a new front-of-pack nutrition label in France: The five-colour Nutri-Score. *Public health panorama*, *03*(04), 712–725. https://iris.who.int/handle/10665/325207

Edenbrandt, A. K., & Smed, S. (2018). Exploring the correlation between self-reported preferences and actual purchases of nutrition labeled products. *Food Policy*, *77*, 71–80. https://doi.org/10.1016/j.foodpol.2018.04.007

Egnell, M., Galan, P., Farpour-Lambert, N. J., Talati, Z., Pettigrew, S., Hercberg, S., & Julia, C. (2020). Compared to other front-of-pack nutrition labels, the Nutri-Score emerged as the most efficient to inform Swiss consumers on the nutritional quality of food products. *PLOS ONE*, *15*(2), Article e0228179. https://doi.org/10.1371/journal.pone.0228179

Ercin, A. E., & Hoekstra, A. Y. (2014). Water footprint scenarios for 2050: A global analysis. *Environment International*, *64*, 71–82. https://doi.org/10.1016/j.envint.2013.11.019

Food and Agriculture Organization of the United Nations. (2015). *2050: Water supplies to dwindle in parts of the world, threatening food security and livelihoods*. Retrieved January 20, 2025, from https://www.fao.org/newsroom/detail/2050-Water-supplies-to-dwindle-in-parts-of-the-world-threatening-food-security-and-livelihoods/en

Food and Agriculture Organization of the United Nations. (n.d.). *Water use*. Retrieved January 20, 2025, from https://www.fao.org/aquastat/en/overview/methodology/water-use

García-González, Á., Achón, M., Carretero Krug, A., Varela-Moreiras, G., & Alonso-Aperte, E. (2020). Food sustainability knowledge and attitudes in the Spanish adult population: A cross-sectional study. *Nutrients*, *12*(10), Article 3154. https://doi.org/10.3390/nu12103154

Gómez-Llanos, E., Durán-Barroso, P., & Robina-Ramírez, R. (2020). Analysis of consumer awareness of sustainable water consumption by the water footprint concept. *Science of The Total Environment*, *721*, Article 137743. https://doi.org/10.1016/j.scitotenv.2020.137743

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Groß, J., Kreis, B. K., Blank, H., & Pachur, T. (2023). Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General*, *152*(11), 3167–3188. https://doi.org/10.1037/xge0001452

Groß, J., Loose, A. M., & Kreis, B. K. (2024). A simple intervention can improve estimates of sugar content. *Journal of Applied Research in Memory and Cognition*, *13*(2), 282–291. https://doi.org/10.1037/mac0000122

Gugenishvili, I., & Laine-Kronberg, A. (2025). The influence of prior knowledge and perceived impact on the connection between attitudes and sustainable behavior. *Sustainable Development*, *33*(3), 4099–4111. https://doi.org/10.1002/sd.3335

Heard, H., & Bogdan, A. (2021). *Healthy and sustainable diets: Consumer poll*. Food Standards Agency. Retrieved from UK Government Web Archive: https://webarchive.nationalarchives.gov.uk/ukgwa/20250404191624/https://www.food.gov.uk/sites/default/files/media/document/healthy-and-sustainable-diets-consumer-poll.pdf

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Hoekstra, A. Y. (2017). Water footprint assessment: Evolvement of a new research field. *Water Resources Management*, *31*(10), 3061–3081. https://doi.org/10.1007/s11269-017-1618-5

Hoekstra, A. Y., Chapagain, A. K., Aldaya, M. M., & Mekonnen, M. M. (2012). *The water footprint assessment manual: Setting the global standard*. Routledge.

Izydorczyk, D., Kreis, B. K., Kilb, M., & Bröder, A. (2025). *# Knowledge Using social media for improving food-related knowledge: A seeding intervention* [Manuscript submitted for publication].

Jeffreys, H. (1961). *The theory of probability* (3rd). Oxford University Press.

Jekauc, D., Voelkle, M. C., Giurgiu, M., & Nigg, C. R. (2024). Unveiling the multidimensional nature of the intention–behavior gap. *European Journal of Health Psychology*, *32*(1), 34–50. https://doi.org/10.1027/2512-8442/a000162

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), Article e0130834. https://doi.org/10.1371/journal.pone.0130834

LaVoie, N. N., Bourne, L. E. J., & Healy, A. F. (2002). Memory seeding: Representations underlying quantitative estimations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1137–1153. https://doi.org/10.1037/0278-7393.28.6.1137

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lohmann, P. M., Gsottbauer, E., Doherty, A., & Kontoleon, A. (2022). Do carbon footprint labels promote climatarian diets? Evidence from a large-scale field experiment. *Journal of Environmental Economics and Management*, *114*, Article 102693. https://doi.org/10.1016/j.jeem.2022.102693

Maher, J. P., & Dunton, G. F. (2020). Within-day time-varying associations between motivation and movement-related behaviors in older adults. *Psychology of Sport and Exercise*, *47*, Article 101522. https://doi.org/10.1016/j.psychsport.2019.04.012

Marghetis, T., Attari, S. Z., & Landy, D. (2019). Simple interventions can correct misperceptions of home energy use. *Nature Energy*, *4*(10), 874–881. https://doi.org/10.1038/s41560-019-0467-2

Mekonnen, M. M., & Gerbens-Leenes, W. (2020). The water footprint of global food production. *Water*, *12*(10), Article 2696. https://doi.org/10.3390/w12102696

Murray, K. B., & Brown, N. R. (2009). A feature-based inference model of numerical estimation: The split-seed effect. *Acta Psychologica*, *131*(3), 221–234. https://doi.org/10.1016/j.actpsy.2009.05.007

Nydrioti, I., & Grigoropoulou, H. (2023). Using the water footprint concept for water use efficiency labelling of consumer products: The Greek experience.

Environmental Science and Pollution Research, *30*(8), 19918–19930. https://doi.org/10.1007/s11356-022-23573-w

Papini, N. M., Chih-Hsiang, Y., Bridgette, D., Tyler B., M., & Lopez, N. V. (2023). External contexts and movement behaviors in ecological momentary assessment studies: A systematic review and future directions. *International Review of Sport and Exercise Psychology*, *16*(1), 337–367. https://doi.org/10.1080/1750984X.2020.1858439

Potter, C., Bastounis, A., Hartmann-Boyce, J., Stewart, C., Frie, K., Tudor, K., Bianchi, F., Cartwright, E., Cook, B., Rayner, M., & Jebb, S. A. (2021). The effects of environmental sustainability labels on selection, purchase, and consumption of food and drink products: A systematic review. *Environment and Behavior*, *53*(8), 891–925. https://doi.org/10.1177/0013916521995473

Pulizzi, F. (2022). *Introducing Nature Water* [Section: News and Opinion, From the Editors]. Retrieved March 20, 2025, from http://sustainabilitycommunity.springernature.com/posts/introducing-nature-water

Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75. https://doi.org/10.1111/tops.12187

Sánchez-Bravo, P., Chambers V, E., Noguera-Artiaga, L., Sendra, E., Chambers IV, E., & Carbonell-Barrachina, Á. A. (2021). Consumer understanding of sustainability concept in agricultural products. *Food Quality and Preference*, *89*, Article 104136. https://doi.org/10.1016/j.foodqual.2020.104136

Sanchez-Sabate, R., & Sabaté, J. (2019). Consumer attitudes towards environmental concerns of meat consumption: A systematic review. *International Journal of Environmental Research and Public Health*, *16*(7), Article 1220. https://doi.org/10.3390/ijerph16071220

Schleyer-Lindenmann, A., Ittner, H., Dauvier, B., & Piolat, M. (2018). Die NEP-Skala – hinter den (deutschen) Kulissen des Umweltbewusstseins. *Diagnostica*, *64*(3), 156–167. https://doi.org/10.1026/0012-1924/a000202

Science Decade. (n.d.). *International decade of sciences for sustainable development (2024-2033)*. Retrieved March 20, 2025, from https://www.un-sciences-decade.org/en

Sheeran, P. (2002). Intention—behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, *12*(1), 1–36. https://doi.org/10.1080/14792772143000003

Sonnenberg, L., Gelsomin, E., Levy, D. E., Riis, J., Barraclough, S., & Thorndike, A. N. (2013). A traffic light food labeling intervention increases consumer awareness of health and healthy choices at the point-of-purchase. *Preventive Medicine*, *57*(4), 253–257. https://doi.org/10.1016/j.ypmed.2013.07.001

Springmann, M., Godfray, H. C. J., Rayner, M., & Scarborough, P. (2016). Analysis and valuation of the health and climate change cobenefits of dietary change. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), 4146–4151. https://doi.org/10.1073/pnas.1523119113

Stan Development Team. (2019). *Stan modeling language: Users guide and reference manual*. https://mc-stan.org

SurveyCircle. (2025). Research website SurveyCircle. Published 2016. Retrieved January 20, 2025, from www.surveycircle.com

Thorndike, A. N., Riis, J., Sonnenberg, L. M., & Levy, D. E. (2014). Traffic-light labels and choice architecture: Promoting healthy food choices. *American Journal of Preventive Medicine*, *46*(2), 143–149. https://doi.org/10.1016/j.amepre.2013.10.002

Truelove, H. B., & Parks, C. (2012). Perceptions of behaviors that cause and mitigate global warming and intentions to perform these behaviors. *Journal of Environmental Psychology*, *32*(3), 246–259. https://doi.org/10.1016/j.jenvp.2012.04.002

United Nations. (2018). *United nations secretary-general's plan: water action decade 2018-2028*. Retrieved January 20, 2025, from

http://www.wateractiondecade.org/wp-content/uploads/2018/03/UN-SG-Action-Plan_Water-Action-Decade-web.pdf

United Nations. (2023). *Summary of proceedings by the president of the general assembly*. Retrieved January 10, 2025, from https://sdgs.un.org/sites/default/files/2023-05/FINAL%20EDITED%20-%20PGA77%20Summary%20for%20Water%20Conference%202023.pdf

United Nations Department of Economic and Social Affairs. (2024). *The Sustainable Development Goals Report 2024 - June 2024*. Retrieved January 23, 2025, from https://desapublications.un.org/publications/sustainable-development-goals-report-2024

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wardle, J., Parmenter, K., & Waller, J. (2000). Nutrition knowledge and food intake. *Appetite*, *34*(3), 269–275. https://doi.org/10.1006/appe.1999.0311

The Water Action Decade is up for review. (2023). *Nature Water*, *1*(3), 209–209. https://doi.org/10.1038/s44221-023-00054-z

Watson, W. L., Kelly, B., Hector, D., Hughes, C., King, L., Crawford, J., Sergeant, J., & Chapman, K. (2014). Can front-of-pack labelling schemes guide healthier food choices? Australian shoppers' responses to seven labelling formats. *Appetite*, *72*, 90–97. https://doi.org/10.1016/j.appet.2013.09.027

Whittall, B., Warwick, S. M., Guy, D. J., & Appleton, K. M. (2023). Public understanding of sustainable diets and changes towards sustainability: A qualitative study in a UK population sample. *Appetite*, *181*, Article 106388. https://doi.org/10.1016/j.appet.2022.106388

Wohldmann, E. L., & Healy, A. F. (2020). Learning and transfer of calorie information. *Applied Cognitive Psychology*, *34*(6), 1485–1494. https://doi.org/10.1002/acp.3727

Appendix A

Materials

Food Items

Table A1

Food Items Used in the Estimation Tasks and the Food-Item Selection Task

Category Food item	Water footprint (L/kg)	Item set	Selection	Label[a]
Vegetables				
Tomatoes	110	A		1
Carrots	130	B		1
Cabbage	240	A		1
Potatoes	248	B		1
Cauliflower	285	A	x	1
Broccoli	285	B	x	1
Zucchini	336	A	x	1
Cucumber	350	B	x	1
Artichokes	818	A		2
Corn	900	B		2
Fruits				
Watermelon	235	A		1
Pineapple	255	B		1
Lemons	360	A		1
Raspberries	413	B		1
Kiwis	514	A	x	2
Oranges	560	B	x	2
Apples	761	B	x	2
Blueberries	845	A	x	2
Bananas	865	B		2
Peaches	910	A		2
Grains and grain products				
Naan bread	1,608	A		3
Sourdough bread	1,608	B		3
Couscous	1,827	B	x	3
Semolina	1,827	A	x	3
Noodles	1,850	A	x	3
Rye flour	1,930	B	x	3
Oat flakes	2,536	B		4
Rice flour	2,628	A		4
Lentils	4,373	A		4
Millet	5,000	B		5

Category Food item	Water footprint (L/kg)	Item set	Selection	Label[a]
Animal-based products				
Milk	1,020	B		3
Crème fraiche	1,898	A		3
Eggs	3,300	B		4
Chicken meat	4,000	A	x	4
Rabbit meat	4,325	B	x	4
Quark	5,060	A	x	5
Feta cheese	5,060	B	x	5
Pork meat	5,360	A		5
Butter	5,553	A		5
Lamb meat	7,850	B		5
Beans, nuts, and seeds				
Soybeans	2,050	B		4
Coconuts	2,500	A		4
Tofu	3,000	A	x	4
Chickpeas	4,177	B	x	4
Pine nuts	5,000	B		5
Kidney beans	5,053	A		5
Macadamia nuts	9,063	B	x	5
Hazelnuts	10,515	A	x	5
Cashew nuts	14,218	B		5
Almonds	15,547	A		5

Note. All water footprints are indicated in liters per kilogram (liters per 1 L for liquids) per food item. A = Item Set A, B = Item Set B. Selection = Whether an item was used in the food-item selection task, investigated in Experiments 2 and 3. Label = Assigned label level from 1 to 5.

[a] The boundaries for assigning food items to a label level were 0–499 L, 500– 999 L, 1,000–1,999 L, 2,000–4,999 L, and more than 5,000 L.

Shopping-list selection task

Table A2

Food Items Used in the Shopping-List Selection Task

Category Food item	Water footprint (L/kg)	List A	List B	List C	List D
Vegetables					
Onions	280	x			
Spinach	292	x			
Mushrooms	322	x		x	x
Pumpkin	353	x	x	x	x
Fruits					
Rhubarb	322		x		
Grapes	608	x	x	x	x
Pears	700		x	x	x
Nectarines	910		x		
Grains and grain products					
Wheat flour	1,849	x	x	x	x
Quinoa	2,654	x	x	x	x
Animal-based products					
Parmesan cheese	4,089	x	x	x	x
Camembert	5,060	x	x	x	
Goat meat	5,500			x	
Beans, nuts, and seeds					
Walnuts	4,918				x
Pecans	9,063	x	x	x	x
Pistachios	11,363				x
Mean water footprint (liters)		2,457	2,560.8	3,019.8	3,591.9

Note. Lists A–D = Four shopping lists, one of which could be selected. An "x" indicates that a food item was included in the corresponding shopping list. The order of the items corresponds to the order in which they were displayed in each shopping list. However, the lists themselves were displayed in random order.

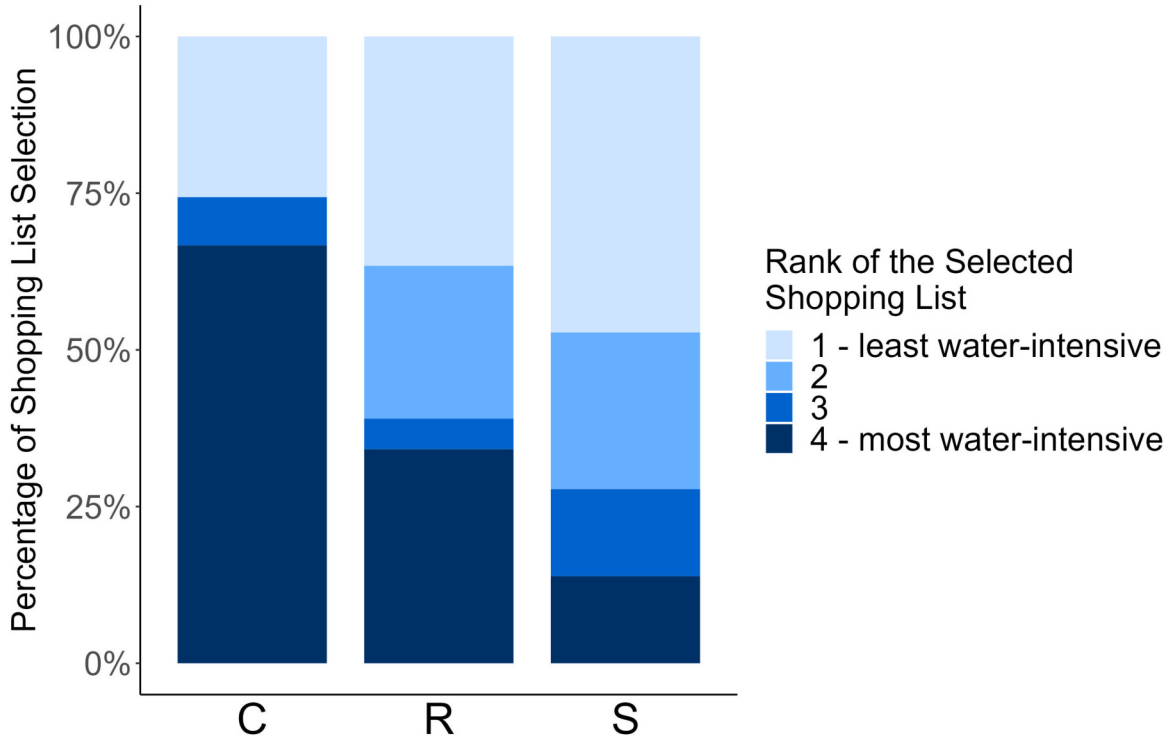
Appendix B

Performance in the Shopping-List Selection Task in Terms of List Rank

To evaluate group differences in participants' performance in the shopping-list selection task we report Bayesian analyses of the mean water footprint of the selected shopping lists in the main text. Here, we report results for the preregistered nonparametric analyses of the rank of the selected shopping lists (see Figure B1). We first applied a Kruskal–Wallis test, revealing significant group differences in mean rank ($\chi^2 = 15$, $p < .001$). Pairwise Dunn–Bonferroni tests showed that the seeding group selected lists with a lower water footprint than the control group ($Z = 3.80$, $p < .001$) and the rule group also outperformed the control group ($Z = 2.45$, $p = .01$). However, no significant difference was found between the seeding and rule groups ($Z = -1.44$, $p = .15$).

Figure B1

Performance in the Shopping-List Selection Task in Experiment 1 in Terms of Shopping-List Rank



Note. The *y* axis shows the percentage of times each shopping-list rank was selected per group. C = Control group. Rule = Rule group. S = Seeding group.

Appendix C

Prior Specification

For all analyses, we specified skeptical priors, centered around 0, indicating that the prior is neutral with respect to potential effects. To assess the sensitivity of our results with respect to different prior specifications, we defined two priors (Priors 1 and 2) for each effect of interest, with Prior 2 being more diffuse, reflecting a broader distribution with less specific assumptions about the magnitude of a possible effect. The results of these sensitivity analyses are reported in Appendix D. To facilitate the specification and intuitive understanding of the intercept priors, we mean-centered the criterion variables for all analyses that examined changes in accuracy (i.e. changes in OME and the Fisher (r-to-z) transformed rank-order correlation ρ between the pre- and post-intervention estimation tasks).

For the analyses of metric knowledge, we defined the following distributions. For all analyses we specified for the intercept parameter a normal distribution `normal(0, 1.5)`, and for the standard deviations of the random effects as well as for the residual standard deviation half-normal distributions, `normal(0, 0.05)` with values > 0. For the analyses of changes in metric knowledge, we further defined for the correlations of the random effects a Lewandowski-Kurowicka-Joe (LKJ) prior distribution with the prior parameter η of 2. Additionally, we defined slope parameters. For the analyses of the main effect of estimation task (pre-intervention estimation task vs. post-intervention estimation task) for each intervention group, we defined a `normal(0, 1)` distribution as Prior 1, and `normal(0, 5)` as Prior 2. For the analyses of the interaction of estimation task and intervention group, we defined `normal(0, 1)` for the slope parameter of the main effects. For the interaction, we defined `normal(0, 1)` for Prior 1 and `normal(0, 5)` for Prior 2.

For the analyses of mapping knowledge, quantified as the Fisher (r-to-z) transformed rank-order correlation ρ, we specified a normal distribution `normal(0, 0.5)` for the intercept parameter. For the standard deviations of the random effects as well as for the residual standard deviation, we defined half-normal distributions,

`normal(0, 0.05)` with values > 0. For the correlations of the random effects we defined an LKJ prior distribution with the prior parameter η of 2. For analyses of the main effect of estimation task, we specified for the slope parameters `normal(0, 0.5)` and `normal(0, 2.5)` as Prior 1 and Prior 2, respectively. For the analyses of the interaction of estimation task and intervention group, we defined `normal(0, 0.5)` for the slope parameter of the main effects. For the interaction, we defined `normal(0, 0.5)` for Prior 1 and `normal(0, 2.5)` for Prior 2.

For the analyses of the selection performance in Experiment 1, we specified a prior of `normal(0, 500)` for the intercept and a half-normal distribution, `normal(0, 300)` with values > 0, for the residual standard deviation. For the slope parameter we specified `normal(0, 500)` for Prior 1 and `normal(0, 2500)` for Prior 2. For Experiment 2 we defined a prior of `normal(0, 1500)` for the intercept and a half-normal distribution, `normal(0, 1000)` with values > 0, for the residual standard deviation. For the slope parameter we specified `normal(0, 1500)` for Prior 1 and `normal(0, 7500)` for Prior 2. For Experiment 3 we defined a prior of `normal(0, 150)` for the intercept and a half-normal distribution, `normal(0, 100)` with values > 0, for the residual standard deviation. For the slope parameter we specified `normal(0, 150)` for Prior 1 and `normal(0, 750)` for Prior 2.

Appendix D

Detailed Results and Sensitivity Analyses

In the main text we report summarized results for the analyses of metric and mapping knowledge with Prior 1. Here, in Tables D1 and D2 we report the detailed results for those analyses. We further report the BFs that resulted from analyses with Prior 2 in Tables D1, D2, and D3. While the different prior specifications slightly affected the size of the BFs, the general conclusions remained the same.

Table D1

Detailed Results of Changes in Metric Knowledge, Quantified as the Order of Magnitude Error, for Prior 1 (and Prior 2)

Item type Effect	Experiment 1 b [95% CI] β BF_{10}	Experiment 2 b [95% CI] β BF_{10}	Experiment 3 b [95% CI] β BF_{10}
Previously estimated items			
ME Seeding	-0.97 $[-1.19, -0.74]$ -1.14 $> 10,000$ ($> 10,000$)	-0.95 $[-1.15, -0.76]$ -1.23 $> 10,000$ ($> 10,000$)	-0.70 $[-0.86, -0.55]$ -1.11 $> 10,000$ ($> 10,000$)
ME Rule	-0.04 $[-0.10, 0.03]$ -0.05 0.06 (0.01)	-	-
ME Label	-	-0.08 $[-0.18, 0.02]$ -0.10 0.17 (0.03)	-0.10 $[-0.18, -0.01]$ -0.13 0.53 (0.10)
ME Control	-0.05 $[-0.09, -0.01]$ -0.06 0.79 (0.15)	-0.05 $[-0.09, -0.02]$ -0.06 1.19 (0.27)	-0.04 $[-0.09, 0.01]$ -0.05 0.08 (0.02)
IA Control*Seeding	-0.93 $[-1.13, -0.72]$ -1.02 $> 10,000$ ($> 10,000$)	-0.90 $[-1.06, -0.73]$ -1.04 $> 10,000$ ($> 10,000$)	-0.67 $[-0.81, -0.52]$ -0.99 $> 10,000$ ($> 10,000$)
IA Control*Rule	0.01 $[-0.05, 0.07]$ 0.02 0.03 (0.01)	-	-
IA Control*Label	-	-0.03 $[-0.11, 0.06]$ -0.03 0.05 (0.01)	-0.06 $[-0.15, 0.03]$ -0.08 0.10 (0.02)
IA Rule*Seeding	0.93 $[-1.14, -0.73]$ -1.08 $> 10,000$ ($> 10,000$)	-	-
IA Label*Seeding	-	-0.87 $[-1.06, -0.68]$ -1.07 $> 10,000$ ($> 10,000$)	-0.61 $[-0.75, -0.46]$ -0.85 $> 10,000$ ($> 10,000$)
New items			
ME Seeding	-0.88 $[-1.11, -0.64]$ -1.06	-0.87 $[-1.07, -0.68]$ -1.16	-0.61 $[-0.76, -0.45]$ -0.99

Item type Effect	Experiment 1 b [95% CI] β BF_{10}	Experiment 2 b [95% CI] β BF_{10}	Experiment 3 b [95% CI] β BF_{10}
	$> 10,000$ ($> 10,000$)	$> 10,000$ ($> 10,000$)	$> 10,000$ ($> 10,000$)
ME Rule	0.02 [−0.10, 0.05] −0.03 0.05 (0.01)	-	-
ME Label	-	−0.07 [−0.16, 0.03] −0.08 0.12 (0.03)	−0.09 [−0.18, 0.01] −0.12 0.24 (0.05)
ME Control	−0.06 [−0.11, −0.02] −0.07 0.83 (0.16)	−0.05 [−0.10, −0.01] −0.06 0.30 (0.06)	−0.06 [−0.11, 0.01] −0.09 0.35 (0.08)
IA Control*Seeding	−0.81 [−1.03, −0.60] −0.92 $> 10,000$ ($> 10,000$)	−0.82 [−0.99, −0.65] −0.97 $> 10,000$ ($> 10,000$)	−0.54 [−0.69, −0.39] −0.82 $> 10,000$ ($> 10,000$)
IA Control*Rule	0.04 [−0.03, 0.11] 0.05 0.07 (0.01)	-	-
IA Control*Label	-	−0.01 [−0.10, 0.07] −0.02 0.05 (0.01)	−0 − 02 [−0.12, 0.08] −0.03 0.06 (0.01)
IA Rule*Seeding	−0.86 [−1.07, −0.64] −1.01 $> 10,000$ ($> 10,000$)	-	-
IA Label*Seeding	-	−0.80 [−0.99, −0.62] −1.00 $> 10,000$ ($> 10,000$)	−0.52 [−0.67, −0.37] −0.73 $> 10,000$ ($> 10,000$)

Note. Reported are the nonstandardized regression coefficients b, their associated credible intervals (CIs), the standardized β coefficients, and the Bayes factors (BFs) for Prior 1. In parentheses we report the BFs for Prior 2. β = Partially standardized regression coefficient, reflecting the difference in Order of Magnitude Error expressed in units of the criterion's standard deviation. As both the group and task variables are categorical, the coefficient was standardized with respect to the criterion. ME = Main effect of estimation task (pre- vs. post-intervention) for each group. IA = Interaction of estimation task (pre- vs. post-intervention) with group. Seeding = Seeding group. Rule = Rule group. Label = Label group. Control = Control group.

Table D2

Detailed Results of Changes in Mapping Knowledge, Quantified as the Fisher (r-to-z) Transformed Rank-Order Correlation ρ, for Prior 1 (and Prior 2)

Item type Effect	Experiment 1 b [95% CI] β BF_{10}	Experiment 2 b [95% CI] β BF_{10}	Experiment 3 b [95% CI] β BF_{10}
Previously estimated items			
ME Seeding	0.70 [0.57, 0.84] 1.38 $> 10,000$ ($> 10,000$)	0.85 [0.71, 0.99] 1.54 $> 10,000$ ($> 10,000$)	0.97 [0.85, 1.11] 1.61 $> 10,000$ ($> 10,000$)
ME Rule	0.63 [0.52, 0.75] 1.40 $> 10,000$ ($> 10,000$)	-	-
ME Label	-	0.91 [0.78, 1.05] 1.57 $> 10,000$ ($> 10,000$)	1.03 [0.90, 1.16] 1.56 $> 10,000$ ($> 10,000$)
ME Control	0.05 [−0.03, 0.12] 0.122 0.13 (0.03)	0.05 [−0.01, 0.12] 0.14 0.30 (0.04)	0.01 [−0.06, 0.09] 0.04 0.08 (0.02)
IA Control*Seeding	0.65 [0.51, 0.80] 1.38 $> 10,000$ ($> 10,000$)	0.80 [0.65, 0.95] 1.55 $> 10,000$ ($> 10,000$)	0.96 [0.80, 1.11] 1.76 $> 10,000$ ($> 10,000$)
IA Control*Rule	0.58 [0.45, 0.71] 1.35 $> 10,000$ ($> 10,000$)	-	-
IA Control*Label	-	0.86 [0.73, 0.99] 1.59 $> 10,000$ ($> 10,000$)	1.01 [0.86, 1.15] 1.72 $> 10,000$ ($> 10,000$)
IA Rule*Seeding	0.09 [−0.08, 0.26] 0.19 0.28 (0.06)	-	-
IA Label*Seeding	-	−0.03 [−0.22, 0.15] −0.06 0.20 (0.04)	−0.03 [−0.20, 0.17] −0.04 0.18 (0.04)
New items			
ME Seeding	0.44 [0.30, 0.57] 1.09 $> 10,000$ ($> 10,000$)	0.65 [0.53, 0.76] 1.48 $> 10,000$ ($> 10,000$)	0.66 [0.53, 0.80] 1.37 $> 10,000$ ($> 10,000$)

Item type Effect	Experiment 1 b [95% CI] β BF_{10}	Experiment 2 b [95% CI] β BF_{10}	Experiment 3 b [95% CI] β BF_{10}
ME Rule	0.67 [0.55, 0.78] 1.45 $> 10,000$ ($> 10,000$)	-	-
ME Label	-	0.59 [0.46, 0.72] 1.28 $> 10,000$ ($> 10,000$)	0.66 [0.54, 0.78] 1.35 $> 10,000$ ($> 10,000$)
ME Control	0.07 [−0.02, 0.16] 0.18 0.31 (0.07)	0.07 [−0.01, 0.15] 0.19 0.33 (0.08)	0.05 [−0.06, 0.15] 0.13 0.13 (0.04)
IA Control*Seeding	0.37 [0.21, 0.52] 0.92 2,546 (923)	0.58 [0.45, 0.71] 1.38 $> 10,000$ ($> 10,000$)	b0.61 [0.44, 0.77] 1.33 $> 10,000$ ($> 10,000$)
IA Control*Rule	0.59 [0.45, 0.74] 1.35 $> 10,000$ ($> 10,000$)	-	-
IA Control*Label	-	0.53 [0.39, 0.67] 1.22 $> 10,000$ ($> 10,000$)	0.60 [0.45, 0.76] 1.31 $> 10,000$ ($> 10,000$)
IA Rule*Seeding	−0.21 [−0.39, −0.04] −0.49 3.27 (0.71)	-	-
IA Label*Seeding	-	0.06 [−0.10, 0.22] 0.12 0.21 (0.04)	0.03 [−0.15, 0.20] 0.06 0.19 (0.04)

Note. Reported are the nonstandardized regression coefficients b, their associated credible intervals (CIs), the standardized β coefficients, and the Bayes factors (BFs) for Prior 1. In parentheses we report the BFs for Prior 2. β = Partially standardized regression coefficient, reflecting the difference in the Fisher (r-to-z) transformed rank-order correlation ρ expressed in units of the criterion's standard deviation. As both the group and task variables are categorical, the coefficient was standardized with respect to the criterion. ME = Main effect of estimation task (pre- vs. post-intervention) for each group. IA = Interaction of estimation task (pre- vs. post-intervention) with group. Seeding = Seeding group. Rule = Rule group. Label = Label group. Control = Control group.

Table D3

Results for the Selection Task for Prior 1 (and Prior 2)

Effect	Experiment 1 BF_{10}	Experiment 2 BF_{10}	Experiment 3 BF_{10}
ME Seeding vs. Control	$> 10,000$ (4,335)	$> 10,000$ ($> 10,000$)	$> 10,000$ (8,906)
ME Rule vs. Control	22 (6)	-	-
ME Label vs. Control	-	$> 10,000$ ($> 10,000$)	2030 (509)
ME Seeding vs. Rule	0.80 (0.17)	-	-
ME Seeding vs. Label	-	0.05 (0.01)	0.09 (0.02)

Note. In parentheses we report the Bayes factors (BFs) for Prior 2. ME = Main effect of group for each reported group comparison. Seeding = Seeding group. Rule = Rule group. Label = Label group. Control = Control group.

Appendix E

Self-Reports of Prior Contact With the Topic and Environmental Mindset

We examined whether participants' self-reported contact with the topic, interest in sustainability, and proenvironmental attitude were associated with initial metric and mapping knowledge, improvements in metric and mapping knowledge through interventions, and performance in the selection tasks. The specific analyses, including priors and model comparisons, can be found on https://osf.io/9aqbs/.

The detailed results for associations between the self-reports and initial metric and mapping accuracy are reported in Table E. Self-reports were not associated with improvements in either metric or mapping knowledge. That is, self-reports were not associated with changes in metric knowledge in the seeding group (all BF_{10} below 0.35 and all CIs included 0) or with changes in mapping knowledge for the seeding, rule, or label group (all BF_{10} below 0.54 and all CIs included 0). Similarly, there was no substantial association of self-reports with participants' selection performance in the shopping-list and food-item selection tasks. All CIs included 0, and while a few BFs provided weak evidence for an association (BF_{10} between 1 and 3), the majority indicated evidence against an effect.

Similarly, there was no substantial association of self-reports with participants' selection performance in the shopping-list and food-item selection tasks. All CIs included 0, and while a few BFs provided weak evidence for an association (BF_{10} between 1 and 3), the majority indicated evidence against an effect.

Table E1

Association of Self-Reports With Initial Metric and Mapping Knowledge

Self-report	Knowledge aspect β	Experiment 1 b [95% CI] β BF_{10}	Experiment 2 b [95% CI] β BF_{10}	Experiment 3 b [95% CI] β BF_{10}
Contact	Metric	**−0.14**, [**−0.22**, **−0.05**] **−0.20** **32**	**−0.09**, [**−0.17**, **−0.02**] **−0.14** **4**	**−0.08**, [**−0.15**, **−0.01**] **−0.14** **4**
	Mapping	0.04, [−0.01, 0.09] 0.15 1.57	0.03, [−0.01, 0.07] 0.11 0.80	0.05, [0.00, 0.10] 0.18 3
Interest sustainability	Metric	−0.09, [−0.19, 0.01] −0.12 1.64	−0.04, [−0.13, 0.05] −0.05 0.48	−0.03, [−0.11, 0.05] −0.05 0.38
	Mapping	0.04, [−0.01, 0.09] 0.13 1.15	0.03, [−0.01, 0.08] 0.11 0.96	**0.07**, [**0.02**, **0.13**] **0.22** **16**
Proenvironmental	Metric	−0.06, [−0.3, 0.18] −0.03 0.66	−0.02, [−0.20, 0.17] −0.01 0.46	0.01, [−0.15, 0.17] 0.01 0.42
	Mapping	0.12, [0.00, 0.24] 0.14 4	0.05, [−0.04, 0.14] 0.08 0.83	0.09, [−0.02, 0.19] 0.12 2

Note. Reported are the nonstandardized regression coefficients b, their associated credible intervals (CIs), the standardized β coefficients, and the Bayes factors (BFs). β = Fully standardized regression coefficient, reflecting the change in Order of Magnitude Error (for metric knowledge) or the Fisher (r-to-z) transformed rank-order correlation ρ (for mapping knowledge), expressed in units of the criterion's standard deviation per one standard deviation increase in the respective self-report questionnaire. Results printed in **bold** indicate unambiguous evidence for an association, where both the Bayes factor (BF) and credible interval (CI) indicate that the respective self-report is associated with initial knowledge. Contact = Prior contact with the topic. Interest sustainability = Interest in sustainability. Proenvironmental = Proenvironmental attitude. Metric = Metric knowledge in the pre-intervention estimation task, quantified as the order of magnitude error. Mapping = Mapping knowledge in the pre-intervention estimation task, quantified as the Fisher (r-to-z) transformed rank-order correlation ρ.

NOTE REGARDING MANUSCRIPT III

At the time this dissertation was submitted, the following manuscript

MANUSCRIPT III

Kreis, B. K., Hermann, A., Pachur, T. & Groß, J. (2025). *Hindsight bias through knowledge updating: A conceptual replication of Groß et al. (2023).* Manuscript invited for revision at *Collabra: Psychology.*

was under review. It has since been published and can be found under the following reference:

Kreis, B. K., Hermann, A., Pachur, T. & Groß, J. (2025). Hindsight bias through knowledge updating: A conceptual replication of Groß et al. (2023). *Collabra: Psychology 11*(1), Article 147499. https://doi.org/10.1525/collabra.147499.

Hindsight Bias Through Knowledge Updating: A Conceptual Replication of Groß et al. (2023)

Barbara K. Kreis[1], Antje Hermann[1], Thorsten Pachur[2,3], and Julia Groß[1]

[1]Department of Psychology, University of Mannheim, Mannheim, Germany

[2]School of Management, Technical University of Munich, Munich, Germany

[3]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Author Note

Correspondence concerning this article should be addressed to Barbara K. Kreis, Experimental Psychology Lab, School of Social Sciences, University of Mannheim, L13, 17, Room 512, D-68161 Mannheim, Germany. E-mail: barbara.kreis@uni-mannheim.de

Abstract

Hindsight bias is the phenomenon that after learning facts about previously judged objects people tend to recall their previous judgments of the objects as closer to the facts than they actually were. Groß, Kreis, Blank, and Pachur (2023) found that hindsight bias emerged not only when people learned actual values for previously judged objects, but also when they learned the values of other objects in the same domain. Moreover, hindsight bias co-occurred with improved estimates for new objects. These findings challenge the traditional view that hindsight bias reflects a cognitive error and instead suggest that it results from adaptive knowledge updating. Groß et al. (2023) obtained their findings in the domain of country populations, a domain where people tend to be unfamiliar with the content and the numerical range (up to several million); this lack of familiarity may affect the link between knowledge updating and hindsight bias. In a high-powered conceptual replication ($N = 300$), we tested whether the findings generalize to the sugar content of food items—a domain where people are more familiar with both content and numerical range (up to 50 grams). Participants provided original judgments for items, learned numerical information, then recalled their original judgments, and lastly provided judgments for a new set of items. Our results replicate the key results of Groß et al. (2023), showing a close link between hindsight bias and knowledge updating in a more familiar domain. We discuss implications for theories of hindsight bias and propose directions for future research.

Keywords: hindsight bias, knowledge updating, replication, real-world estimation

Introduction

Have you ever wondered about the amount of sugar in a jam doughnut? Say you guess 15 grams, then look it up and discover that the true figure is 29 grams. Later, when recalling your initial estimate, you might confidently think, "I guessed 25 grams." What you have experienced is *hindsight bias* (Blank et al., 2007; Fischhoff, 1975, 2025; Hawkins & Hastie, 1990; Roese & Vohs, 2012). Hindsight bias occurs when people estimate a quantity or predict an event's outcome, learn the actual value or result, and then recall their initial judgment as being closer to the truth than it actually was. This phenomenon has been demonstrated across various contexts and materials, including real-world knowledge (e.g., answers to almanac questions, numerical estimates; Erdfelder & Buchner, 1998; Groß et al., 2023) and outcomes (e.g., medical cases, elections; Arkes et al., 1988; Blank et al., 2003). Hindsight bias is traditionally viewed as an illusion or cognitive error that results from the limitations of the human mind (e.g., Pohl, 2022). One account posits that it is due to people's use of the *anchoring-and-adjustment* heuristic (Tversky & Kahneman, 1974), the assumption being that people use the new information (i.e., the actual value or outcome) as an anchor when reconstructing their initial judgment, but fail to adjust sufficiently away from it (Hawkins & Hastie, 1990; Tversky & Kahneman, 1974).

More recently, an alternative view has been proposed, according to which hindsight bias might represent a by-product of an adaptive learning process—knowledge updating (Groß et al., 2023; Hoffrage et al., 2000; Nestler et al., 2012). From this perspective, people update the knowledge base they used to generate their initial judgment based on the new information. When asked to recall their initial judgment, they then rejudge the object based on this updated knowledge, leading to a systematic discrepancy from the initial judgment (Hoffrage et al., 2000). From this *knowledge-updating-and-rejudgment* perspective, hindsight bias is a consequence of adaptive information integration, rather than reflecting a cognitive error.

A set of studies by Groß et al. (2023) has supported the knowledge-updating account of hindsight bias. The authors derived a set of predictions (detailed below) and

tested them in the context of numerical estimation, with country populations as the knowledge domain. Their results demonstrated for the first time that hindsight bias arises not only when people learn the actual values of objects (as commonly shown), but also when they learn the values of other objects in the same domain. That is, hindsight bias emerged when participants were presented with any type of numerical information that allowed for knowledge updating. These results challenge the traditional theoretical assumptions on the psychological underpinnings of hindsight bias.

In light of their novelty and theoretical relevance, it is important to test whether Groß et al.'s (2023) findings can be replicated and extended beyond the domain of country populations. This extension is critical because people have limited familiarity with both the objects of this domain (Brown & Siegler, 1993; LaVoie et al., 2002) and the numerical range it covers (in the millions; Landy et al., 2013; Resnick et al., 2017), which may affect both hindsight bias and knowledge updating processes. Why? Greater familiarity with objects is known to facilitate memory formation and recall (Bellana et al., 2021) and could thus reduce hindsight bias (e.g., Christensen-Szalanski & Willham, 1991; Hertwig et al., 2003). Additionally, participants' calibration to a given numerical range, which is typically poorer in less familiar ranges, is linked to errors in their numerical judgments (e.g., Boyce-Jacino et al., 2022; Landy et al., 2013; Louie et al., 2015; Resnick et al., 2017); it might thus also influence how effectively participants improve their judgments through knowledge updating. However, it is unclear whether hindsight bias and knowledge updating are equally influenced by familiarity and whether the degree of familiarity impacts the observed association between them. It is thus important to explore whether the link established by Groß et al. (2023) generalizes to a domain with which people are more familiar and that involves a smaller and thus more familiar numerical range. In this article, we address this issue by conducting a preregistered (https://osf.io/2pr76) conceptual replication of Groß et al. (2023) for another domain, namely, the sugar content of food items.

The Knowledge-Updating Account of Hindsight Bias

To examine the knowledge-updating account of hindsight bias, Groß et al. (2023) connected hindsight bias research to research on *seeding effects* in numerical real-world estimation (Brown & Siegler, 1993). This research has shown that presenting actual values for a set of objects, so-called *seed facts* (e.g., that Spain has a population of 48 million inhabitants), leads to an improvement of people's estimation accuracy not only for subsequent judgments of the seeded objects (e.g., Spain), but also for nonseeded objects from the same domain (e.g., France or Canada; Bröder et al., 2023; Brown & Siegler, 1993; Groß et al., 2024; LaVoie et al., 2002). We refer to this effect as *transfer learning*. It is assumed that the seed facts lead to an update of the underlying metric knowledge about the domain (i.e., knowledge of the range, central tendency, or distribution of country populations; Brown & Siegler, 1993), thus improving estimation accuracy.

Groß et al. (2023) proposed that the actual values of objects that lead to hindsight bias should also function as seed facts, prompting an update of the underlying metric knowledge and thereby improving subsequent judgments. Incorporating insights from seeding research, the authors derived three predictions that should hold if hindsight bias emerges, at least partly, due to knowledge updating. First, presenting actual values for a set of objects should not only elicit hindsight bias for those objects, but also prompt an update of the underlying metric knowledge for the whole domain, leading to improved estimates for both the original objects and other objects from the domain (i.e., transfer learning). Second, if hindsight bias is a consequence of knowledge updating, then any numerical information that people can use to update their domain knowledge should induce hindsight bias. Therefore, presenting values for other objects from the same domain should not only elicit transfer learning but also trigger hindsight bias for the previously estimated objects (hindsight bias via domain information). Third, when the values presented do not provide information relevant to updating the knowledge base (e.g., because they refer to a different domain), no hindsight bias should be triggered (no hindsight bias via irrelevant information).

To investigate hindsight bias and transfer learning simultaneously and test whether they might co-occur as consequences of knowledge updating, Groß et al. (2023) developed an *Integrated Hindsight-Bias-and-Seeding Paradigm*. In this paradigm, participants first judge a set of objects, are then presented with numerical information, are subsequently asked to recall their initial judgments (as a measure of hindsight bias), and finally provide estimates for other objects from the same domain (as a measure of transfer learning). Using this paradigm, Groß et al. (2023, Exp. 2) obtained support for all three predictions. The knowledge domain used in Groß et al. (2023), country populations, is commonly used in seeding research. Can their results be replicated in a domain where people are arguably more familiar with the content and operate with a more familiar numerical range?

The Present Study

Here we conduct a conceptual replication of Groß et al. (2023, Exp. 2), using the sugar content of food items as the knowledge domain. It is plausible that people are relatively familiar with this domain, as they are often exposed to information on sugar content—for example, via nutrition labels or recipes. Relative familiarity with the numerical range of the objects can also be assumed (typically 0 to 50 grams of sugar per serving or purchase size).

As in the original study, we used the Integrated Hindsight-Bias-and-Seeding Paradigm (Groß et al., 2023). Participants first completed an estimation task for a set of objects (original judgments, OJs). In an information phase, they were then presented with the actual values for those same objects (feedback group), the values for other objects in the same domain (domain-information group), or the actual values for the original objects but relabeled as belonging to another domain (irrelevant-information group). In the control group, no actual values were presented. Instead, participants read a text unrelated to the topic of the study. Participants were then asked to recall their original estimates (recall of original judgments, ROJs) and to provide estimates for a new set of objects (new judgments, NJs). If the ROJs are closer to the actual values than the OJs, then this would suggest the presence of hindsight effects; if the NJs are

closer to the actual values than the OJs, then this would suggest the presence of transfer learning effects.

We tested the three predictions that Groß et al. (2023, Exp. 2) derived from the knowledge-updating account of hindsight bias. Our hypotheses were as follows. First, we expected that presenting the actual sugar content values for a set of previously estimated food items would both trigger hindsight effects for those food items and improve estimation accuracy for new food items (hindsight and transfer learning effects in the feedback group). Second, we expected that not only presenting actual sugar content values for a set of previously estimated food items but also presenting values for other items from the same domain would also both trigger hindsight effects for the previously estimated items and improve estimation accuracy for new food items (hindsight and transfer learning effects in the domain-information group). Third, we expected that presenting the actual sugar content values but relabeled as referring to an unrelated domain—here, longitudes of European cities—would not lead to either hindsight effects or improved estimation accuracy for new food items (no hindsight or transfer learning effects in the irrelevant-information group). Finally, we did not expect to observe hindsight or transfer learning effects in the control group, where no actual values were presented (no hindsight or transfer learning effects in the control group).

Method

The experiment was preregistered (see https://osf.io/2pr76).

Participants

The experiment was conducted online. Participants were recruited through Prolific (www.prolific.co) with the same general eligibility criteria as in the original study by Groß et al. (2023, Exp. 2): Participation was limited to native German speakers aged 18 to 45 years. Additionally, we required that the study was conducted on a laptop, desktop computer, or tablet, rather than a smartphone. This was to ensure that the items, which were accompanied by food images, were properly displayed. We additionally excluded individuals who reported type 1 or type 2 diabetes or an eating disorder (e.g., anorexia nervosa, bulimia nervosa, binge-eating disorder). The median

completion time of the experiment was 18.09 minutes, and participants received a fixed compensation of £4.80.

Overall, $N = 340$ participants completed the experiment. They were, on average, 27.9 years old, with 142 identifying as women, 193 as men, and 5 as diverse. In the original study ($N = 295$), participants were, on average, 27.8 years, with 116 women, 178 men, and 1 other. In the present study, most participants reported a university degree as their highest level of education (157), followed by a certificate of higher secondary education (122), vocational training (24), a certificate of lower secondary education (24), and a PhD (13). The majority of participants were currently working (177) or studying (university or college; 126); 13 participants were not working but looking for work; 4 were not working and not looking for work; 14 were apprentices; 6 were in high school. In the original sample, 130 participants were working, 122 studying (university or college), 29 looking for work, 7 in high school, 6 apprentices, and 1 was retired. The present sample was thus similar to that of Groß et al. (2023, Exp. 2) in terms of these demographic characteristics.

The target sample size for the present study was determined by conducting an a priori frequentist simulation-based power analysis with the `mixedpower` package in R (Kumle et al., 2021). We based the simulation on the results of the effect of interest of Experiment 2 in Groß et al. (2023): the hindsight effect in the domain-information group. To account for uncertainty of the expected effect size, we simulated power for the smaller boundary (in absolute terms) of the $CI_{90\%}$ of the effect. We ran 5,000 simulations and defined a critical t value of 2, reflecting an α level of 5%. The simulations showed that a power of 80% would be achieved with 80 participants, resulting in an overall target sample size of 320 for four groups. To account for potential exclusions, we collected data from 20 additional participants, resulting in a total sample size of 340.

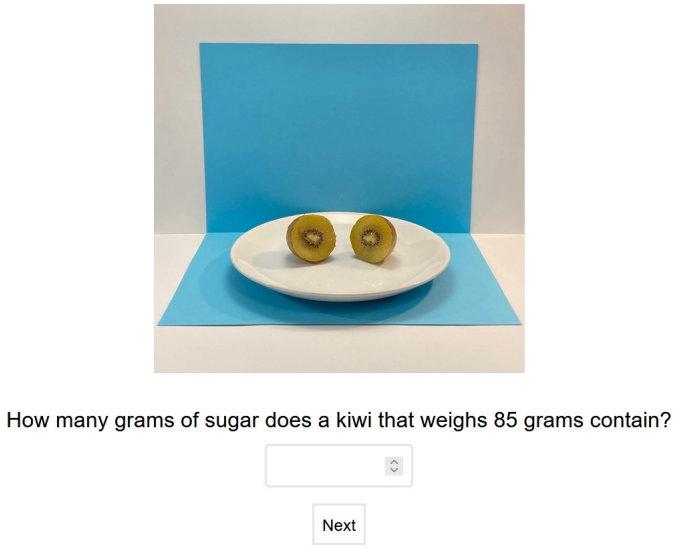
Materials

Participants estimated the sugar content (in grams) of common food items. In contrast to the original study, where items were presented as text only (e.g., “How many

people live in Peru?"), we included an image and specified the weight of each food item (e.g., "How many grams of sugar are in a kiwi that weighs 85 grams?") to clarify the size of the food item and increase ecological validity (see Figure 1 for a sample item).

Figure 1

Sample Item from the Original Judgment Task



Note. Translated from German.

Alt text. Example screen showing an image of a halved kiwi fruit on a plate, placed against a blue backdrop. Below the image, a question prompts participants to estimate the sugar content of a 85-g kiwi, with a response field and a "Next" button.

As in the original study by Groß et al. (2023), we used 96 items. These were taken from eight food subcategories (grains and grain products; vegetables; fruit; milk and dairy products; fish, meat, and sausage; oils and fats; drinks; and sweets and snacks), with each subcategory containing 12 items.[1] For most items, participants were asked to estimate the sugar content of a typical serving size, such as a banana; for others, the sugar content of a typical purchase size, such as a can of coconut milk. The food items varied in sugar content ($M = 7.6$ grams, $SD = 7.7$, $min = 0$, $max = 33$) and degree of processing (based on the NOVA classification system; Monteiro et al., 2019):

[1] The definition of the eight food subcategories was based on the categorization guidelines by the German Society for Nutrition (Deutsche Gesellschaft für Ernährung e. V., 2024) and the Federal Center for Nutrition (Bundeszentrum für Ernährung, 2023).

unprocessed or minimally processed foods (e.g., a banana or whole milk), processed culinary ingredients (e.g., a tablespoon of butter or a tablespoon of honey), processed foods (e.g., a serving of spaghetti or a can of sweetcorn), and ultra-processed foods (e.g., a bowl of sausage salad or a glass of energy drink). We photographed all items and edited the pictures to blur any visible labels, nutritional information, and numbers. All food images are available at https://osf.io/hr5bv/. A list of all food items, their subcategories, size (serving or purchase size in grams or milliliters), and sugar content (in grams) can be found in Appendix A1. From these 96 items, we created three sets of 32 items each. Each set contained four items from each of the eight food subcategories (see Appendix A2 for details). The sets were designed to be comparable in terms of degree of processing and statistical characteristics of the sugar content (mean, median, range). Thus, assignment of items to item sets followed similar principles as in the original study.

In the experiment, one set of items was presented to all groups in both the OJ and ROJ task. In the information phase, actual values for the same set were presented to participants in the feedback group; the same values, but relabeled as longitudes of European cities (which have a similar numerical range to the sugar content values), to participants in the irrelevant-information group. Participants in the domain-information group were presented with values for a second set of food items. Finally, all groups were presented with a third set of food times in the NJ task.

To assess participants' familiarity with the sugar content of food items in terms of engagement, we asked the following question: "At the beginning of this study, you answered several questions about the sugar content of food items. How *frequently*, prior to this study, have you engaged with this topic (e.g., at school, at work, or in your leisure time)?" (translated from German). Participants indicated their answer on a 7-point scale, ranging from *very rarely* to *very frequently*.

Design and Procedure

The design and procedure of the present experiment closely followed that of Groß et al. (2023, Exp. 2). The experiment was programmed with lab.js (Henninger

et al., 2022) and hosted via Jatos (Lange et al., 2015). It consisted of four phases: Participants (1) completed the OJ task, (2) were presented with actual values of different types depending on the experimental group to which they were randomly assigned, (3) completed the ROJ task, and (4) completed the NJ task. This resulted in a three (task: OJ, ROJ, NJ) by four (group: control, feedback, domain information, irrelevant information) mixed design.

After providing informed consent, participants entered the first phase, the OJ task, where they estimated the sugar content of 32 food items (e.g., "How many grams of sugar does a kiwi that weighs 85 grams contain?"). The items were presented sequentially in a randomized order for each participant. Participants took a median of 6.4 seconds per estimate (which was faster than in the original study: 9.4 seconds). In the second phase, the information phase, participants were presented with different types of values depending on their assigned group. Those in the *feedback* group were shown the actual values of the 32 food items from the OJ task (e.g., "A kiwi that weighs 85 grams contains 8 grams of sugar."). Those in the *domain-information* group were shown the values of a different set of 32 food items (e.g., "A plum that weighs 75 grams contains 7 grams of sugar."). Those in the *irrelevant-information* group saw the same values as participants in the feedback group (i.e., the actual sugar content values) but relabeled as the longitudes of 32 well-known European cities, along with the flag of the respective country (e.g., "Basel, a major city in Switzerland, is located at longitude 8."). This domain was selected to match the mean and range of the sugar content values across the three sets (see Appendix A2 for details). In all three groups, each item was presented for 8 seconds, resulting in an overall presentation duration of 256 seconds. In the *control* group, participants were presented with a text that was unrelated to the topic of the study (filler task) for an overall duration of 256 seconds. The presentation time was extended from 5 seconds per item in the original study to 8 seconds per item here to account for the more complex combination of quantity information and an accompanying image.

In the third phase, the ROJ task, participants were asked to recall their original

judgments in the OJ task for all 32 items, presented in the same randomized order (e.g. "How many grams of sugar does a kiwi that weighs 85 grams contain? What was your ORIGINAL answer?"). In the fourth and final phase, the NJ task, participants estimated the sugar content of a new set of 32 food items ("How many grams of sugar does a tangerine that weighs 90 grams contain?"). The assignment of sets was counterbalanced, with each set being presented with comparable frequency across participants and experimental groups. After completing all four phases, participants were asked whether they had cheated or clicked through the experiment and whether they had any problems with the display of images, before indicating their prior engagement with the domain. Finally, they had the opportunity to enter any additional comments into an open response field.

Data Diagnostics

We preregistered several data assessment steps comparable to those applied in Groß et al. (2023, Exp. 2) to ensure data quality by excluding both extreme outliers and data indicating non-compliance. First, Prolific automatically excluded participants who did not finish the experiment within the specified 87-minute time limit. Second, we excluded all participants whose meta-data indicated that contrary to the eligibility criteria they had used a smartphone (six participants). Third, we excluded participants who indicated non-compliance, such as cheating (one participant) or just clicking through the experiment (one participant). Fourth, we excluded participants who reported technical problems or disruptions (two participants who reloaded the experiment, one participant who reported a disturbance that led to a long break during the experiment, and three participants who reported problems with the display of several food images). Fifth, we excluded one participant for providing estimates that exceeded the weight of the food item for 10 or more items, as such sugar content is not possible. Sixth, we planned to exclude cases in which participants provided estimates in under 1000 ms for 10 or more items, as such short estimation processes are implausible; however, there were no such cases. Seventh, we excluded 21 participants whose estimates had a median absolute deviation from the actual values (see Eq. 1 below)

that exceeded the threefold interquartile range in at least one of the three experimental tasks (three in the control group, five in the feedback group, eight in the domain-information group, and five in the irrelevant-information group). Eighth, we excluded individual estimates that exceeded the weight of the food item (30 responses, 0.09% of all responses) and responses given in less than 1000 ms (one response, 0.003% of all responses). The final sample consisted of $N = 300$ participants. Sample size was comparable across experimental groups, with $n = 76$ in the control group, $n = 73$ in the feedback group, $n = 74$ in the domain-information group, and $n = 77$ in the irrelevant-information group.

Analytic Approach

To quantify estimation accuracy in each task—OJ, ROJ, and NJ—we calculated $|\Delta|$, the absolute deviation of the estimate from the actual value, for each item i and for each participant j:

$$|\Delta_{ij}| = |estimate_{ij} - actual_i| \tag{1}$$

Smaller values of $|\Delta|$ indicate a higher estimation accuracy. In Groß et al.'s (2023) original study, estimation accuracy was quantified with a logarithmic deviation measure, the order of magnitude error (OME). The use of this measure was not indicated in the present study for two reasons. First, the OME is indicated for highly skewed distributions spanning several orders of magnitude. This applied to the country population domain used by Groß et al. (2023), but does not apply to sugar content. Second, the frequent occurrence of zero values in sugar content estimates (8.9% of all estimates) and actual values (19.9% of all actual values) makes logarithmic measures problematic, as the logarithm of zero is undefined. In all other respects, the analytic approach in the present study was identical to that in the original study.

We applied Bayesian linear mixed-effects models to compare estimation accuracy across judgment tasks and groups, defining $|\Delta|$ as the criterion variable. As fixed effects, we specified the task, information group, and their interaction. Furthermore, we included random intercepts and random slopes for participants and items to capture

by-person and by-item variability in estimation accuracy. For parameter estimation we used the `brms` package (Bürkner, 2017, 2018), which calls `STAN` for MCMC sampling (Stan Development Team, 2019). We report prior specification and sensitivity analyses in Appendix B. The general conclusions were robust across two alternative specifications of the priors.

To test our hypotheses, we compared a full model including the fixed-effect predictor of interest, M_1, with a baseline model M_0, that did not include that predictor but did include all random effects present in the full model.[2] The model comparisons are described in detail in "Results". We compared the models using the `bayes_factor` function in `brms`, which computes Bayes Factors (BF) based on bridge sampling (e.g., Gronau et al., 2017). The BF_{10} indicates the evidence for the alternative hypothesis relative to the null hypothesis, when comparing the full model M_1 to the baseline model M_0.[3] The data and the analysis code are available at https://osf.io/hr5bv/.

We performed additional analyses investigating another facet of domain knowledge—namely, the accuracy of the ranking of objects (mapping knowledge; Brown & Siegler, 1993)—as detailed in Appendix C.

Treatment of Perfectly Accurate Judgments

For all analyses, we excluded all OJs and corresponding ROJs where the OJ was an exact match for the actual value (7.4%), as in these cases neither a hindsight effect nor a transfer learning effect can occur. (As NJ items are new items, there were no corresponding NJs to be excluded.) In Groß et al. (2023, Exp. 2), there were no cases of

---

[2] We preregistered all analyses for the fixed effect task, comparing estimation accuracy in the OJ and ROJ/NJ tasks separately by group. All additional interaction analyses were conducted to gain further insights into the relational dynamics between the effects.

[3] A BF_{10} below 1/10 is generally taken to indicate strong evidence, a BF_{10} between 1/10 and 1/3 moderate evidence, and a BF_{10} between 1/3 and 1 weak evidence for M_0. A BF_{10} larger than 10 is generally taken to indicate strong evidence, a BF_{10} between 3 and 10 moderate evidence, and a BF_{10} between 1 and 3 weak evidence for M_1 (e.g., Jeffreys, 1998; van Doorn et al., 2023).

perfectly accurate judgments.[4]

Treatment of Perfect OJ Reproductions

As in Groß et al. (2023, Exp. 2), we also identified perfect OJ reproductions, that is, cases where the ROJ matched the OJ exactly (e.g., due to correct recollection of the OJ or successful guessing). In such cases, hindsight effects are zero, and the higher the percentage of such cases, the smaller the overall size of hindsight effects will be. Any difference in hindsight effects between groups might thus be due to a genuine difference in hindsight effects, to different percentages of perfect OJ reproductions, or to both. As in previous investigations (Dehn & Erdfelder, 1998; Erdfelder & Buchner, 1998; Erdfelder et al., 2007), the percentages of perfect OJ reproductions in our study differed significantly between groups ($X^2(3) = 34.04$, $p < .001$), being lower in the feedback (35.23%) and the domain-information groups (33.08%) than in the irrelevant-information (39.67%) and control groups (39.67%). Following previous studies (e.g., Erdfelder & Buchner, 1998; Groß & Pachur, 2019; Pohl, 2007), we therefore excluded all perfectly matching OJ/ROJ pairs from our analyses.[5] In Groß et al. (2023, Exp. 2), the percentage of such cases was similar across groups (and they were therefore not excluded); additional analyses showed that the conclusions were unaffected by whether perfect OJ reproductions were excluded or not.

Results

Overall, the results of the present study replicate those of the original study (Groß et al., 2023, Exp. 2), with few exceptions. Unless noted otherwise, the observed patterns of results are identical. Participants' self-reported engagement with the sugar

---

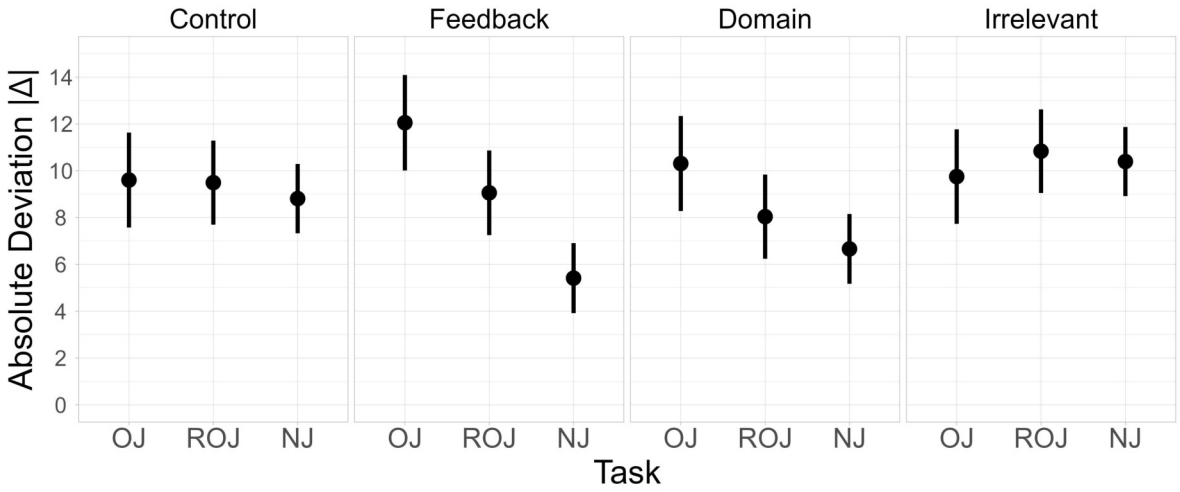
[4] We decided to exclude these cases due to the substantial number of perfectly accurate judgments in the present study, although we had not preregistered this step. Robustness analyses showed that the patterns of results remained the same, whether these cases were excluded or not.

[5] We preregistered excluding OJ = ROJ pairs (if the percentages of such cases differed between groups) for the hindsight effect analyses. We decided to also exclude them from the transfer learning analyses for better comparability. Robustness analyses showed that the transfer learning results were unaffected by whether those cases were excluded or not.

content of food items (ranging from 1 = *very rarely* to 7 = *very frequently*) was, on average, 3.44 ($SD = 1.45$).[6] Figure 2 plots estimation accuracy for the four groups and three judgment tasks.

Figure 2

Estimation Accuracy



Note. Shown are the conditional predictions based on the mixed-effects model (estimated means and 95% credible intervals). OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment. Control = Control group, Feedback = Feedback group, Domain = Domain-information group, Irrelevant = Irrelevant-information group.

Alt text. Four-panel plot comparing estimation accuracy across experimental groups: control, feedback, domain-information, and irrelevant-information. Each panel displays estimation accuracy (absolute deviation $|\Delta|$) on the y-axis across three tasks on the x-axis—OJ, ROJ, and NJ—with points and vertical error bars..

---

[6] We tested whether self-reported engagement with the domain was associated with initial estimation accuracy (i.e., estimation accuracy in the OJ task), and the results were ambiguous. The credible interval of the regression weight did not include zero ($b = -0.51$, $CI_{95\%} = [-0.95, -0.07]$), indicating that more engagement was associated with higher initial estimation accuracy; however, the BF indicated moderate evidence against an effect ($BF_{10} = 0.24$). In addition, self-reported engagement was not associated with either hindsight effects or transfer learning effects in any of the experimental groups (all credible intervals included zero and all BF_{10} were below 0.32).

Transfer Learning Effects

We first tested whether presenting actual values led to transfer learning effects. This would be the case if $|\Delta|$ (i.e., the deviation of the estimate from the actual value) was smaller for the NJs than for the OJs. To test for transfer learning effects, we compared a model including the fixed-effect predictor task (OJ versus NJ) to a baseline model that did not include the predictor. As expected, in both experimental groups that saw domain-relevant values, there were transfer learning effects (feedback group: $b = -6.59$, $CI_{95\%} = [-8.63, -4.55]$, $\text{BF}_{10} > 10,000$; domain-information group: $b = -3.71$, $CI_{95\%} = [-5.09, -2.32]$, $\text{BF}_{10} > 10,000$). An interaction analysis indicated that the transfer learning effect was larger in the feedback group than in the domain-information group ($b = -3.11$, $CI_{95\%} = [-5.06, -1.16]$, $\text{BF}_{10} = 11$). In contrast, in Groß et al. (2023, Exp. 2), the size of the effect did not differ between the two groups. There was no transfer learning effect in the irrelevant-information group ($b = 0.52$, $CI_{95\%} = [-0.45, 1.49]$, $\text{BF}_{10} = 0.08$) or the control group ($b = -0.92$, $CI_{95\%} = [-2.00, 0.15]$, $\text{BF}_{10} = 0.19$), with an interaction analysis indicating no difference between these two groups ($b = 1.35$, $CI_{95\%} = [-0.06, 2.77]$, $\text{BF}_{10} = 0.37$).

Hindsight Effects

Next we tested whether presenting actual values triggered a hindsight effect. This would be the case if $|\Delta|$ was smaller for the ROJs than for the OJs. To test for hindsight effects, we compared a model including the fixed-effect predictor task (OJ versus ROJ) to a baseline model that did not include the predictor. As expected, there was a hindsight effect in the feedback group ($b = -2.88$, $CI_{95\%} = [-4.19, -1.54]$, $\text{BF}_{10} = 268$). Importantly, there was also a hindsight effect in the domain-information group ($b = -2.25$, $CI_{95\%} = [-3.32, -1.16]$, $\text{BF}_{10} = 80$). An interaction analysis showed that the size of the hindsight effect did not differ between these two groups ($b = -0.70$, $CI_{95\%} = [-2.11, 0.71]$, $\text{BF}_{10} = 0.07$). Critically, there was no hindsight effect in the irrelevant-information group ($b = 0.87$, $CI_{95\%} = [-0.06, 1.81]$, $\text{BF}_{10} = 0.23$) or the control group ($b = 1.10$, $CI_{95\%} = [-0.15, 2.35]$, $\text{BF}_{10} = 0.25$), with an interaction analysis indicating no difference between the control and the irrelevant-information

group ($b = -0.26$, $CI_{95\%} = [-1.15, 0.64]$, $\mathrm{BF}_{10} = 0.05$). In other words, when the actual values were presented to participants but relabeled as referring to a different domain (as in the irrelevant-information group), this did not impact the ROJs. This analysis was based on $|\Delta|$, the absolute deviation of the estimate from the actual value of each item. It is possible that the relabeled values may have affected ROJs in other ways—for instance, via the central tendency of the presented values, or only the most recently presented values. Such results could point to the operation of anchoring-and-adjustment processes. However, additional analyses showed no indication for such alternative influences of the relabeled values (for details, see Appendix D, and Groß et al., 2023).

In sum, the results replicate the key novel finding of Groß et al. (2023, Exp. 2). Hindsight effects were triggered for objects even when participants were presented with the values of other objects in the same domain, not only when they were presented with the actual values of the original objects. That is, hindsight effects occurred when knowledge updating was enabled through the provision of relevant numerical information. Thus, hindsight bias also seems to be triggered by knowledge updating in a domain where people are relatively familiar with both the content and the underlying numerical range. Our results also replicate the finding in Groß et al. (2023) that a hindsight effect is not triggered when the numerical information presented is relabeled as referring to a different domain, thus hindering knowledge updating. Together, these findings support a clear link between hindsight bias and knowledge updating.

Discussion

Hindsight bias has long been seen as resulting from limitations of the human mind (Hawkins & Hastie, 1990; Tversky & Kahneman, 1974). More recently, it has been proposed that hindsight bias might in fact be a by-product of an adaptive learning process—knowledge updating (Groß et al., 2023; Hoffrage et al., 2000; Nestler et al., 2012). Investigating hindsight bias in the context of country population estimates, Groß et al. (2023) provided evidence that hindsight bias for objects can be elicited not only by actual values for those specific objects, but also by values for other objects in the same domain. This finding suggests that hindsight judgments reflected re-judgments of

the objects based on an updated metric representation of the knowledge domain.

In the present study, we tested whether these findings can be replicated in the domain of sugar content of food items, with which participants are likely to be more familiar. There are several indicators for such higher familiarity. First, the initial accuracy of the ranking of objects, quantified as the rank-order correlation between estimates and actual values in the OJ task, was higher for sugar content ($\rho = 0.61$) than for country populations ($\rho = 0.52$; Groß et al., 2023).[7] Second, perfectly accurate judgments—that is, cases where the OJ matched the actual value exactly—were more frequent for sugar content than for country populations (7.4% vs. 0%; Groß et al., 2023), OJs were given faster (6.4 vs. 9.4 seconds per response; Groß et al., 2023), and participants reported more prior engagement with sugar content than did a comparable sample for country populations ($M = 3.44$ vs. $M = 2.58$; Kreis et al., 2024).[8]

As hypothesized, we observed indications for a close link between hindsight bias and knowledge updating, even in this more familiar judgment domain. Most importantly, we replicated the key novel finding by Groß et al. (2023) that hindsight effects for previously estimated objects can be triggered even without presenting the actual values of those objects—it can suffice to present the actual values of other objects in the same domain. Further, presenting actual values for previously estimated objects but relabeled as referring to an unrelated domain did not lead to either hindsight effects or transfer learning effects. Hindsight bias thus consistently emerged when participants were presented with relevant numerical information that allowed for knowledge updating, but not when that numerical information was relabeled, thus hindering knowledge updating.

It should be noted that our finding that a robust link between hindsight bias and

---

[7] For further information on the analyses and results of the rank-order correlations, see Appendix C.

[8] A direct comparison between the domains in terms of initial estimation accuracy is not possible because it needed to be quantified differently: For country populations, a logarithmic accuracy measure was needed to take into account the skewness and large range of values; the same did not apply to sugar content, where logarithmic measures were unsuitable due to the high prevalence of zero values (as discussed in "Analytic Approach").

knowledge updating emerges even in a domain with which people are more familiar does not imply that the amount of prior knowledge is irrelevant for hindsight bias and knowledge updating. Both are likely to be affected in tandem by whether prior knowledge is high or low. The greater the amount of prior knowledge, the less knowledge updating there is likely to be and hence the smaller hindsight bias. To examine this link for the present study, we compared the size of transfer learning and hindsight effects across the eight food subcategories (see section "Materials"), and examined whether the sizes of the two effects covaried with the amount of prior knowledge of the subcategories, reflected in participants' initial estimation accuracy (i.e., accuracy of the OJs). As Figure 3 shows, the sizes of transfer learning and hindsight effects seem to be closely linked and both depend on the amount of prior knowledge in a given subcategory. Food subcategories with higher initial estimation accuracy showed smaller transfer learning and hindsight effects, while those with lower initial estimation accuracy showed greater transfer learning and hindsight effects. For example, in the feedback group, for the Oils and fats subcategory, where initial estimation accuracy was rather high ($|\Delta| = 7.08$), transfer learning and hindsight effects were relatively small, at $|\Delta| = 4.20$ and 1.93, respectively; whereas for the Drinks subcategory, where initial estimation accuracy was rather low ($|\Delta| = 20.26$), transfer learning and hindsight effects were relatively large, at $|\Delta| = 9.46$ and 5.08, respectively.[9]

Implications

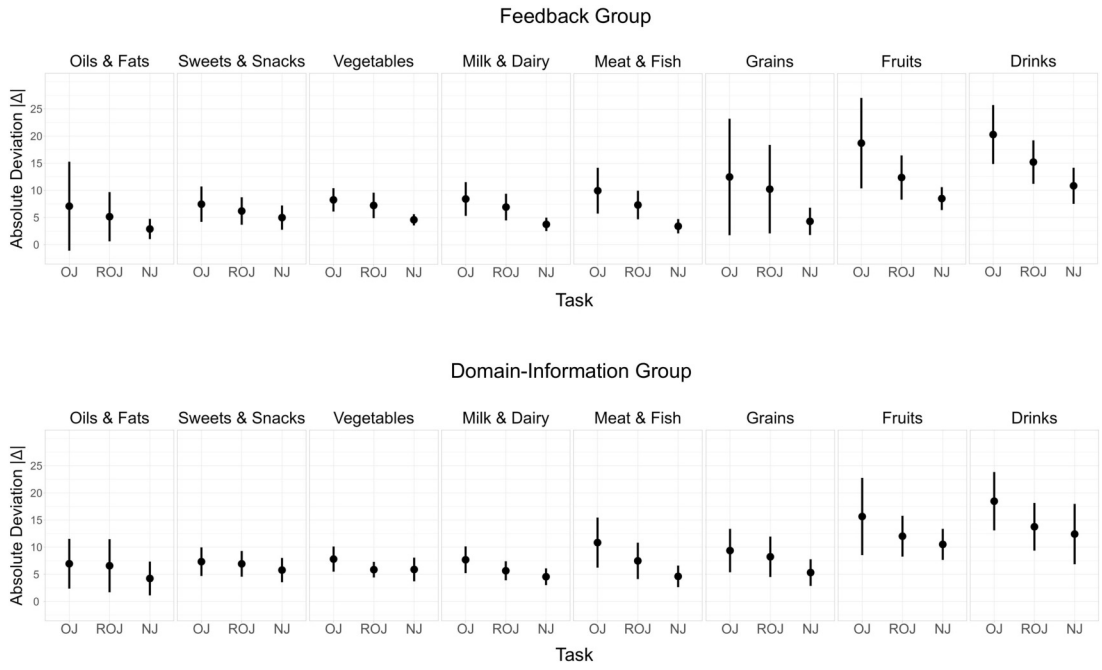
The results of our conceptual replication provide further evidence that hindsight bias, while traditionally viewed as cognitive error, might actually reflect knowledge updating and rejudgment (Groß et al., 2023; Hoffrage et al., 2000; Nestler et al., 2012). The findings indicate that hindsight bias is triggered whenever people learn numerical information about a domain that they can use to update their underlying knowledge—whether actual values for previously estimated items or values for other items from the same domain.

This finding highlights the pervasiveness of hindsight bias and may help explain

[9] Reported are the conditional predictions of the means based on the mixed-effects model.

Figure 3

Comparison of Estimation Accuracy by Food Subcategory in the Feedback and Domain-Information Groups



Note. Shown are the conditional predictions based on the mixed-effects models by group and food subcategory (estimated means and 95% credible intervals). Food subcategories are sorted ascending by initial estimation accuracy (estimated mean $|\Delta|$ in the OJ task) in the feedback group. OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment.

Alt text. Two-panel plot comparing estimation accuracy across eight food subcategories for two experimental groups: feedback (top) and domain-information (bottom). Each panel contains eight subplots, one per food category, showing estimation accuracy (absolute deviation $|\Delta|$) on the y-axis across three tasks on the x-axis—OJ, ROJ, and NJ—using points with vertical error bars.

why it has proven so difficult to eliminate in the context of numerical estimation. For example, hindsight bias has been shown to persist when participants are fully informed about the bias and then retested (Pohl & Hell, 1996). Even discrediting the actual values prior to the ROJ task (Erdfelder & Buchner, 1998) or presenting them as another person's estimates (Pohl, 1998) seems to only partially reduce hindsight bias. Yet this robustness of hindsight bias is to be expected if people integrate any numerical

information that can be used to inform their underlying metric knowledge about a domain rather automatically (Fischhoff, 1975; Hawkins & Hastie, 1990). To the extent that people usually rejudge an object when trying to reconstruct their initial judgment for the object, hindsight bias thus represents an inevitable by-product of knowledge updating.

However, the results of the present study indicate that processes other than knowledge updating might also contribute to hindsight bias. In both the feedback and the domain-information groups, hindsight effects were smaller than transfer learning effects, with the ROJs being further away from the actual values than the NJs (see also Groß et al., 2023, Exp. 2). If both types of responses were exclusively (re-)judgments based on updated knowledge, that should not be the case. However, note that in the ROJ task, participants are asked to recall their original judgments, whereas in the NJ task, they are asked to provide estimates for new objects. Thus, participants seem to engage in additional processes in the ROJ task (e.g., adjusting the ROJ using episodic memory traces of the OJ to account for inaccuracies in initial estimation).

Outlook

While our findings add to the increasing evidence for a knowledge-updating account of hindsight bias, future research could further test this account by addressing additional aspects. For example, another approach to investigating the link between knowledge updating and hindsight bias could be to manipulate the time interval between the information phase and the ROJ task. Research on seeding effects suggests that the updating of metric knowledge is maintained over extended periods, with transfer learning effects remaining stable for up to four months regardless of whether the specific seed facts can still be recalled (Brown & Siegler, 1996; LaVoie et al., 2002). Thus, to the extent that hindsight bias is driven by processes of knowledge updating, it should remain stable over similar time frames. If hindsight bias diminishes over time, this would suggest that additional processes—such as anchoring and adjustment, which relies more on the recall of specific values—may be involved.

In the present study, we concluded that hindsight bias in the context of

real-world estimation is primarily driven by knowledge updating, and we found no evidence for the influence of anchoring-and-adjustment processes—which would have manifested as a hindsight effect in the irrelevant-information group. It is conceivable, however, that the contribution of anchoring-and-adjustment processes depends on the order of presentation of actual values and ROJ task. In the present study, all actual values were presented first, and only then did participants provide ROJs. If actual values and ROJs were presented and provided alternately, anchoring processes might play a larger role; this is because the relabeled values could be more directly linked to subsequent ROJs, potentially increasing their influence (Hawkins & Hastie, 1990; Tversky & Kahneman, 1974).

Conclusion

The nature and underlying mechanisms of hindsight bias have been debated for decades (Blank et al., 2007; Fischhoff, 1975; Hawkins & Hastie, 1990; Hoffrage et al., 2000; Roese & Vohs, 2012). The present study, together with that of Groß et al. (2023), contributes to the increasing evidence that hindsight bias reflects knowledge updating processes rather than a cognitive error (see also Hoffrage et al., 2000; Nestler et al., 2012). With this replication, which demonstrates the robustness and generalizability of the link between hindsight bias and knowledge updating across domains, we contribute to establishing a solid empirical foundation that can inform the continued development of theory and future research on hindsight bias (Eronen & Bringmann, 2021).

CRediT authorship contribution statement

Barbara K. Kreis served as lead for investigation, data curation, formal analysis, visualization, writing-original draft and validation and in a supporting role for resources and software. Antje Hermann served as a lead for resources and software and in a supporting role for investigation, data curation and validation. Thorsten Pachur served in a supporting role for conceptualization and supervision. Julia Groß served in a supporting role for validation, resources and writing-original draft. Barbara K. Kreis, Antje Hermann, Thorsten Pachur and Julia Groß served equally in writing-review and editing. Barbara K. Kreis, Antje Hermann and Julia Groß served equally in conceptualization and methodology. Barbara K. Kreis and Julia Groß served equally in project administration and supervision. Thorsten Pachur and Julia Groß served equally in funding acquisition.

Acknowledgements

The authors thank Susannah Goss for editing the manuscript.

Funding

This work was funded by Grant GR-4649/4-1 (PA 1925/2-1) from the German Research Foundation (DFG) which was awarded to Julia Groß and Thorsten Pachur. Barbara K. Kreis was further supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

Data accessibility statement

Data and analysis code and research materials are available via the Open Science Framework https://osf.io/hr5bv/.

Ethics statement

The study protocol was approved by the Ethics Committee of the University of Mannheim.

Competing interests

The authors have no conflicts of interest to disclose.

References

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, *73*(2), 305–307. https://doi.org/10.1037/0021-9010.73.2.305

Bellana, B., Mansour, R., Ladyka-Wojcik, N., Grady, C., & Moscovitch, M. (2021). The influence of prior knowledge on the formation of detailed and durable memories. *Journal of Memory and Language*, *121*, 104264. https://doi.org/10.1016/j.jml.2021.104264

Blank, H., Fischer, V., & Erdfelder, E. (2003). Hindsight bias in political elections. *Memory*, *11*(4–5), 491–504. https://doi.org/10.1080/09658210244000513

Blank, H., Musch, J., & Pohl, R. F. (2007). Hindsight bias: On being wise after the event. *Social Cognition*, *25*(1), 1–9. https://doi.org/10.1521/soco.2007.25.1.1

Boyce-Jacino, C., Peters, E., Galvani, A. P., & Chapman, G. B. (2022). Large numbers cause magnitude neglect: The case of government expenditures. *Proceedings of the National Academy of Sciences*, *119*(28), e2203037119. https://doi.org/10.1073/pnas.2203037119

Bröder, A., Dülz, E., Heidecke, D., Wehler, A., & Weimann, F. (2023). Improving carbon footprint estimates of food items with a simple seeding procedure. *Applied Cognitive Psychology*, *37*(3), 651–659. https://doi.org/10.1002/acp.4060

Brown, N. R. (2002, January). Real-world estimation: Estimation modes and seeding effects. In *Psychology of learning and motivation* (pp. 321–359, Vol. 41). Academic Press. https://doi.org/10.1016/S0079-7421(02)80011-1

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534. https://doi.org/10.1037/0033-295X.100.3.511

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review*, *3*(3), 385–388. https://doi.org/10.3758/BF03210766

Bundeszentrum für Ernährung. (2023). Die Ernährungspyramide [The food pyramid]. Retrieved August 5, 2024, from https://www.bzfe.de/ernaehrung/die-ernaehrungspyramide/die-ernaehrungspyramide-eine-fuer-alle/

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*(1), 147–168. https://doi.org/10.1016/0749-5978(91)90010-Q

Dehn, D. M., & Erdfelder, E. (1998). What kind of bias is hindsight bias? *Psychological Research*, *61*(2), 135–146. https://doi.org/10.1007/s004260050020

Deutsche Gesellschaft für Ernährung e. V. (2024). DGE-Ernährungskreis [DGE Nutrition Circle]. Retrieved August 5, 2024, from http://www.dge.de/gesunde-ernaehrung/gut-essen-und-trinken/dge-ernaehrungskreis/

Erdfelder, E., Brandt, M., & Bröder, A. (2007). Recollection biases in hindsight judgments. *Social Cognition*, *25*(1), 114–131. https://doi.org/10.1521/soco.2007.25.1.114

Erdfelder, E., & Buchner, A. (1998). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 387–414. https://doi.org/10.1037/0278-7393.24.2.387

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fischhoff, B. (2025). Fifty years of hindsight bias research-Reflection on Fischhoff (1975). *Journal of Experimental Psychology. Human Perception and Performance*, *51*(2), 143–150. https://doi.org/10.1037/xhp0001232

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Groß, J., Kreis, B. K., Blank, H., & Pachur, T. (2023). Knowledge updating in real-world estimation: Connecting hindsight bias and seeding effects. *Journal of Experimental Psychology: General*, *152*(11), 3167–3188. https://doi.org/10.1037/xge0001452

Groß, J., Loose, A. M., & Kreis, B. K. (2024). A simple intervention can improve estimates of sugar content. *Journal of Applied Research in Memory and Cognition*, *13*(2), 282–291. https://doi.org/10.1037/mac0000122

Groß, J., & Pachur, T. (2019). Age differences in hindsight bias: A meta-analysis. *Psychology and Aging*, *34*(2), 294–310. https://doi.org/10.1037/pag0000329

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*(3), 311–327. https://doi.org/10.1037/0033-2909.107.3.311

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. https://doi.org/10.3758/s13428-019-01283-5

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory*, *11*(4-5), 357–377. https://doi.org/10.1080/09658210244000595

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 566–581. https://doi.org/10.1037/0278-7393.26.3.566

Jeffreys, H. (1998, August). *The theory of probability*. Oxford University Press.

Kreis, B. K., Groß, J., & Pachur, T. (2024). Real-world estimation taps into basic numeric abilities. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02575-4

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. *Cognitive Science*, *37*(5), 775–799. https://doi.org/10.1111/cogs.12028

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

LaVoie, N. N., Bourne, L. E. J., & Healy, A. F. (2002). Memory seeding: Representations underlying quantitative estimations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1137–1153. https://doi.org/10.1037/0278-7393.28.6.1137

Louie, K., Glimcher, P. W., & Webb, R. (2015). Adaptive neural coding: From biological to behavioral decision-making. *Current Opinion in Behavioral Sciences*, *5*, 91–99. https://doi.org/10.1016/j.cobeha.2015.08.008

Monteiro, C. A., Cannon, G., Lawrence, M., Costa Louzada, M., & Pereira Machado, P. (2019). *Ultra-processed foods, diet quality, and health using the NOVA classification system.* Food and Agriculture Organization of the United Nations. https://openknowledge.fao.org/server/api/core/bitstreams/5277b379-0acb-4d97-a6a3-602774104629/content

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology*, *103*(4), 689–717. https://doi.org/10.1037/a0029461

Pohl, R. F. (2007). Ways to assess hindsight bias. *Social Cognition*, *25*(1), 14–31. https://doi.org/10.1521/soco.2007.25.1.14

Pohl, R. F. (2022). *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* [3rd ed.]. Routledge.

Pohl, R. F. (1998). The effects of feedback source and plausibility of hindsight bias. *European Journal of Cognitive Psychology*, *10*(2), 191–212. https://doi.org/10.1080/713752272

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*(1), 49–58. https://doi.org/10.1006/obhd.1996.0064

Resnick, I., Newcombe, N. S., & Shipley, T. F. (2017). Dealing with big numbers: Representation and understanding of magnitudes outside of human experience. *Cognitive Science*, *41*(4), 1020–1041. https://doi.org/10.1111/cogs.12388

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2023). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*, *28*(6), 1404–1426. https://doi.org/10.1037/met0000472.supp

Stan Development Team. (2019). Stan modeling language: Users guide and reference manual. //mc-stan.org/

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2023). Bayes factors for mixed models. *Computational Brain & Behavior*, *6*, 1–13. https://doi.org/10.1007/s42113-021-00113-2

Appendix A

Materials

Table A1

Food Items Used in the Experiment

Subcategory Food item	Serving/Purchase size (g or ml)	Sugar content (in g)	Item set
Grains and grain products			
1 soft pretzel	90	7	A
1 potato	110	1	A
1 pre-packaged pizza dough	400	4	A
1 rusk	10	1	A
1 croissant	70	4	B
1 slice of toast	25	1	B
1 serving of cooked spaghetti	125	4	B
1 wrap	65	4	B
1 slice of farmhouse bread	55	1	C
1 sweet potato	320	13	C
1 serving of cooked rice	125	0	C
1 bread roll	60	1	C
Vegetables			
1 can of corn	140	11	A
1 jar of artichokes	165	1	A
1 kohlrabi	265	10	A
1 onion	90	6	A
1 jar of bell peppers	165	9	B
1 jar of white asparagus	115	2	B
1 cucumber	420	8	B
1 carrot	60	4	B
1 can of kidney beans	265	9	C
1 jar of pickled gherkins	185	8	C
1 bell pepper	150	8	C
1 zucchini	225	4	C

Subcategory Food item	Serving/Purchase size (g or ml)	Sugar content (in g)	Item set
Fruit			
1 can of pineapple slices	275	33	A
1 kiwi	85	8	A
1 orange	170	13	A
1 serving of dried apple slices	30	17	A
1 pear	205	18	B
1 can of peach halves	250	28	B
1 tangerine	90	8	B
1 serving of raisins	30	20	B
1 banana	110	17	C
5 dates	40	22	C
1 can of apricot halves	240	18	C
1 plum	75	7	C
Milk and dairy products			
1 cup of vegan coconut yogurt	160	2	A
1 tablespoon of cream cheese	20	1	A
1 glass of full fat milk	200	9	A
1 mozzarella	125	1	A
1 cup of low-fat quark cheese	250	9	B
1 glass of buttermilk	200	8	B
1 oven-baked cheese	180	1	B
1 slice of butter cheese	30	0	B
1 cup of plain yogurt	150	7	C
1 package of feta cheese	200	0	C
1 glass of oat milk	200	10	C
1 slice of Gouda cheese	30	0	C
Fish, meat, and sausage			
1 serving of smoked salmon	25	0	A
1 serving of sausage salad	230	6	A
1 slice of cooked ham	20	0	A
8 vegetarian chicken nuggets	180	0	A
1 can of tuna	80	0	B
1 Fleischkaese[a]	220	1	B
4 German meat patties	200	3	B
1 vegetarian sausage	50	0	B
5 fish fingers	150	1	C
1 slice of salami	10	0	C
2 vegetarian cordon bleu	200	1	C
2 Wiener sausages	100	1	C

Subcategory Food item	Serving/Purchase size (g or ml)	Sugar content (in g)	Item set
Oils and fats			
1 avocado	150	1	A
1 can of coconut milk	400	8	A
1 tablespoon of butter	20	0	A
1 serving of almonds	30	1	A
1 cup of sour cream	200	7	B
1 can of black olives	85	0	B
1 tablespoon of mayonnaise	15	0	B
1 serving of pistachios	30	2	B
1 cup of cream	200	6	C
1 tablespoon of hummus	20	0	C
1 tablespoon of sunflower oil	5	0	C
1 serving of peanuts	30	2	C
Drinks			
1 glass of energy drink	200	22	A
1 glass of carrot juice	200	12	A
1 glass of rhubarb spritzer	200	14	A
1 glass of lemonade	200	16	A
1 glass of apple spritzer	200	13	B
1 glass of coke	200	21	B
1 glass of peach iced tea	200	10	B
1 glass of orange juice	200	18	B
1 glass of bitter lemon	200	24	C
1 glass of ginger ale	200	18	C
1 glass of tomato juice	200	6	C
1 glass of grape spritzer	200	16	C
Sweets and snacks			
1 cup of vanilla pudding	150	17	A
1 tablespoon of strawberry jam	20	11	A
1 serving of coated peanuts	30	2	A
1 serving of sweet popcorn	30	8	A
1 tablespoon of honey	20	15	B
1 serving of crisps	30	1	B
1 chocolate doughnut	60	8	B
1 serving of gummy bears	30	14	B
1 cookie	20	6	C
1 tablespoon of nougat hazelnut spread	20	8	C
1 raspberry jam doughnut	95	29	C
1 serving of pretzels	30	0	C

Note. [a] German meat loaf specialty.

Table A2

Sets of Foods (Sugar Content in Brackets) and Set of Cities (Longitude in Brackets) Used in the Experiment

Set A		Set B		Set C		City Set[a]	
8 vegetarian chicken nuggets	(0)	1 vegetarian sausage	(0)	1 slice of salami	(0)	Bordeaux	(−1)
1 serving of smoked salmon	(0)	1 can of tuna	(0)	1 slice of Gouda cheese	(0)	Alicante	(0)
1 slice of cooked ham	(0)	1 slice of butter cheese	(0)	1 package of feta cheese	(0)	London	(0)
1 tablespoon of butter	(0)	1 can of black olives	(0)	1 serving of cooked rice	(0)	Le Havre	(0)
1 jar of artichokes	(1)	1 tablespoon of mayonnaise	(0)	1 tablespoon of hummus	(0)	Cambridge	(0)
1 tablespoon of cream cheese	(1)	1 piece of Fleischkaese	(1)	1 tablespoon of sunflower oil	(0)	Rouen	(0)
1 mozzarella	(1)	1 oven-baked cheese	(1)	1 serving of pretzels	(0)	Tarragona	(1)
1 rusk	(1)	1 serving of crisps	(1)	2 vegetarian cordon bleu	(1)	Norwich	(1)
1 potato	(1)	1 slice of toast	(1)	5 fish fingers	(1)	Ibiza	(1)
1 avocado	(1)	1 jar of white asparagus	(2)	2 Wiener sausages	(1)	Toulouse	(1)
1 serving of almonds	(1)	1 serving of pistachios	(2)	1 bread roll	(1)	Andorra la Vella	(2)
1 cup of vegan coconut yogurt	(2)	4 German meat patties	(3)	1 slice of farmhouse bread	(1)	Barcelona	(2)
1 serving of coated peanuts	(2)	1 wrap	(4)	1 serving of peanuts	(2)	Lille	(3)
1 pre-packaged pizza dough	(4)	1 carrot	(4)	1 zucchini	(4)	Brussels	(4)
1 onion	(6)	1 croissant	(4)	1 cup of cream	(6)	Bergen	(5)
1 serving of sausage salad	(6)	1 serving of cooked spaghetti	(4)	1 glass of tomato juice	(6)	Marseilles	(5)
1 soft pretzel	(7)	1 cup of sour cream	(7)	1 cookie	(6)	Cologne	(7)
1 kiwi	(8)	1 tangerine	(8)	1 plum	(7)	Basel	(8)
1 can of coconut milk	(8)	1 glass of buttermilk	(8)	1 cup of plain yogurt	(7)	Turin	(8)
1 serving of sweet popcorn	(8)	1 chocolate doughnut	(8)	1 jar of pickled gherkins	(8)	Zurich	(8)
1 glass of full fat milk	(9)	1 cucumber	(8)	1 bell pepper	(8)	Genoa	(9)
1 kohlrabi	(10)	1 cup of low-fat quark cheese	(9)	1 tablespoon of nougat hazelnut spread	(8)	Milan	(9)
1 can of corn	(11)	1 jar of bell peppers	(9)	1 can of kidney beans	(9)	Hamburg	(10)
1 tablespoon of strawberry jam	(11)	1 glass of peach iced tea	(10)	1 glass of oat milk	(10)	Oslo	(11)
1 glass of carrot juice	(12)	1 glass of apple spritzer	(13)	1 sweet potato	(13)	Salzburg	(13)
1 orange	(13)	1 serving of gummy bears	(14)	1 glass of grape spritzer	(16)	Prague	(14)
1 glass of rhubarb spritzer	(14)	1 tablespoon of honey	(15)	1 banana	(17)	Zagreb	(16)
1 glass of lemonade	(16)	1 pear	(18)	1 can of apricot halves	(18)	Bratislava	(17)
1 serving of dried apple slices	(17)	1 glass of orange juice	(18)	1 glass of ginger ale	(18)	Stockholm	(18)
1 cup of vanilla pudding	(17)	1 serving of raisins	(20)	5 dates	(22)	Budapest	(19)
1 glass of energy drink	(22)	1 glass of coke	(21)	1 glass of bitter lemon	(24)	Thessaloniki	(23)
1 can of pineapple slices	(33)	1 can of peach halves	(28)	1 raspberry jam doughnut	(29)	Kiev	(31)

Note. [a] Cities were selected such that their longitudes approximately matched the mean and range of sugar contents across the three sets. The values presented to the participants in the irrelevant-information group (i.e., sugar content relabeled as longitudes) therefore only approximated the actual longitudes.

Appendix B

Prior Specification and Sensitivity Analyses

As recommended by Schad et al. (2023), we verified that our prior specifications produced plausible data by running prior predictive checks. Further, to facilitate specification of the priors, we mean-centered the criterion variable $|\Delta|$. (Note that we omitted to preregister our intention to mean-center the criterion variable.)

We specified the following prior distributions. For the intercept parameter, we specified a normal distribution `normal(0,15)`. For the slope parameters, we defined two different priors to examine how these different specifications would affect the results of the analysis. One was "skeptical" with regard to a potential effect, placing a lot of prior probability around zero, `normal(0,10)`. The other prior was "weakly informative," `normal(-5,10)`, with more probability mass around an effect with a negative sign, indicating that the deviation $|\Delta|$ decreases from the OJ task to the other tasks (ROJ and NJ), but still with probability mass around zero allowing for unanticipated effects in the opposite direction. For the standard deviations of the random effects (i.e., participants, items), we specified half-normal distributions, `normal(0,15)` with values > 0 as priors. For the correlation of the random effects, we defined a Lewandowski-Kurowicka-Joe (LKJ) prior distribution with the prior parameter η of 2. For the residual standard deviation, we specified a half-normal distribution, `normal(0,15)` with values > 0.

For comparison, we report the results for both the "weakly informative" prior (reported in the main text) and the "skeptical" prior in Table B1. As shown, while the different priors led to slightly different Bayes Factors, the general conclusions and result patterns remained the same.

Table B1

Results with Skeptical Prior [and Weakly Informative Prior] on the Slope Parameter

	Control	Feedback	Domain	Irrelevant	IA Control × Irrelevant	IA Feedback × Domain
	BF_{10}	BF_{10}	BF_{10}	BF_{10}	BF_{10}	BF_{10}
Transfer learning effect	0.20 [0.19]	>10,000 [>10,000]	>10,000 [>10,000]	0.08 [0.08]	0.42 [0.37]	13 [11]
Hindsight effect	0.06 [0.05]	280 [268]	76 [80]	0.25 [0.23]	0.29 [0.25]	0.11 [0.07]

Note. Control = Control group, Feedback = Feedback group, Domain = Domain-information group, Irrelevant = Irrelevant-information group. IA = Interaction. OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment.

Appendix C
Mapping Knowledge

As in Groß et al. (2023), in addition to metric knowledge, we also report results for another component of numerical domain knowledge, namely, mapping knowledge (i.e., knowledge of the ordering of objects in a domain; Brown & Siegler, 1993, 1996). Mapping knowledge is typically quantified by the rank-order correlation between estimates and actual values.

As in the original study, we expected that there would be no improvement in the rank-order correlation for the NJ relative to the OJ task (i.e., no transfer learning effect), as mapping knowledge usually only improves for the objects for which actual values are presented (Brown, 2002; Brown & Siegler, 1993; Groß et al., 2024). For the same reason, we expected to observe an improvement in the rank-order correlation for the ROJ relative to OJ task (i.e., a hindsight effect) in the feedback group only, as only this group receives actual values for the previously estimated objects.

For the analyses of mapping knowledge, we Fisher (r-to-z) transformed and then mean-centered the rank-order correlation. Unlike in the analyses of metric knowledge, we included cases of perfectly accurate judgment (i.e., OJ = actual value) and perfect OJ reproduction (i.e., ROJ = OJ), as they do not distort the computation of person-level rank-order correlations.

As in the original study, we used Bayesian linear-mixed effects models. We defined the rank-order correlation of each participant j as the criterion of the models. As fixed effects, we specified the task and further included random intercepts and random slopes for participants. We specified the following priors. For the intercept and slope parameters, we used `normal(0,0.5)`; for the standard deviation of the random effects (i.e., intercept and slope for participants) and the residual standard deviation, we used half-normal distributions with values > 0, `normal(0,0.5)`; for the correlations of the random effects, we used an LKJ prior distribution with $\eta = 2$. We compared a model including the fixed-effect predictor task (OJ versus NJ for transfer learning effects; OJ versus ROJ for hindsight effects) to a baseline model without the predictor.

Tables C1 and C2 present the results. The analyses of transfer learning effects yielded evidence for improved mapping knowledge in both the feedback and the domain-information group. While these findings are inconsistent with those of Groß et al. (2023), who did not find any transfer learning effects for mapping knowledge, they are in line with some previous research on seeding effects (Bröder et al., 2023; Brown & Siegler, 1996). As expected, the analyses of hindsight effects showed improved mapping knowledge in the feedback group. In addition, and in line with the observed transfer learning effect in the domain-information group, we also found a hindsight effect in the domain-information group.

While this finding deviates from the findings of Groß et al. (2023), it aligns well with the knowledge-updating account of hindsight bias, which predicts that if information elicits knowledge updating, as indicated by transfer learning effects, hindsight effects should co-occur. As expected, we

found no hindsight or transfer learning effects for mapping knowledge in the control or irrelevant-information groups.

In sum, while some results deviated from the original study, the general pattern of results remained in line with the predictions of the knowledge-updating account of hindsight bias, with transfer learning effects consistently co-occurring with hindsight effects.

Table C1

Rank-Order Correlations Between Estimated and Actual Values

	Control		Feedback		Domain		Irrelevant	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
OJ Task	0.61	(0.13)	0.63	(0.10)	0.60	(0.14)	0.60	(0.14)
ROJ Task	0.62	(0.15)	0.70	(0.11)	0.63	(0.14)	0.60	(0.13)
NJ Task	0.64	(0.13)	0.69	(0.12)	0.65	(0.14)	0.61	(0.14)

Note. Shown are the means and standard deviations of the rank-order correlations by group and task. Control = Control group. Feedback = Feedback group. Domain = Domain-information group. Irrelevant = Irrelevant-information group. OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment.

Table C2

Improvements in Mapping Knowledge

	Control	Feedback	Domain	Irrelevant
	BF_{10}	BF_{10}	BF_{10}	BF_{10}
Transfer learning effect (OJ vs. NJ)	0.70	3	8	0.05
Hindsight effect (OJ vs. ROJ)	0.08	>10,000	5	0.03

Note. Control = Control group. Feedback = Feedback group. Domain = Domain-information group. Irrelevant = Irrelevant-information group. OJ = Original Judgment, ROJ = Recall of Original Judgment, NJ = New Judgment.

Appendix D

Investigating Anchoring Effects in the Irrelevant-Information Group

In the main article, we reported analyses examining whether presenting actual values framed as irrelevant information (i.e., longitudes of European cities) elicited a hindsight effect. From the finding that no such effect occurred, we concluded that hindsight judgments were not influenced by anchoring-and-adjustment processes. This conclusion is based on the assumption that processes of anchoring and adjustment might be elicited by the presentation of the full set of relabeled actual values. However, it is possible that exposure to the presented values might have influenced the ROJs via anchoring-and-adjustment processes in other ways. Specifically, participants could have been influenced by the central tendency (i.e., the mean or median) of the presented values, or by the last items presented. To explore these possibilities, we calculated the mean and median $|\Delta|$—the absolute difference between each estimate and the mean or median of the actual values presented during the information phase—for both the OJ and the ROJ task.[10] We specified the same priors as for the main analyses, and chose the *weakly informative* prior for the slope parameter (see Appendix B). We applied the same analytical steps as in the main analysis; that is, we compared a full model that included the main effect of task to a baseline model that did not. The results indicated that the ROJs were not influenced by either the mean ($BF_{10} = 0.20$) or the median ($BF_{10} = 0.32$) of the presented values. Additionally, we tested for recency effects by calculating the recency mean and median $|\Delta|$ for the last five values presented to each participant. Again, no effects on the ROJs were found, for either the recency mean $|\Delta|$ ($BF_{10} = 0.15$) or the recency median $|\Delta|$ ($BF_{10} = 0.16$).

[10] In the preregistration, we mistakenly stated that we would calculate the log-transformed difference between the estimates and the mean/median of the actual values. For reasons described in the section "Analytic Approach", we did not log transform values in this study.